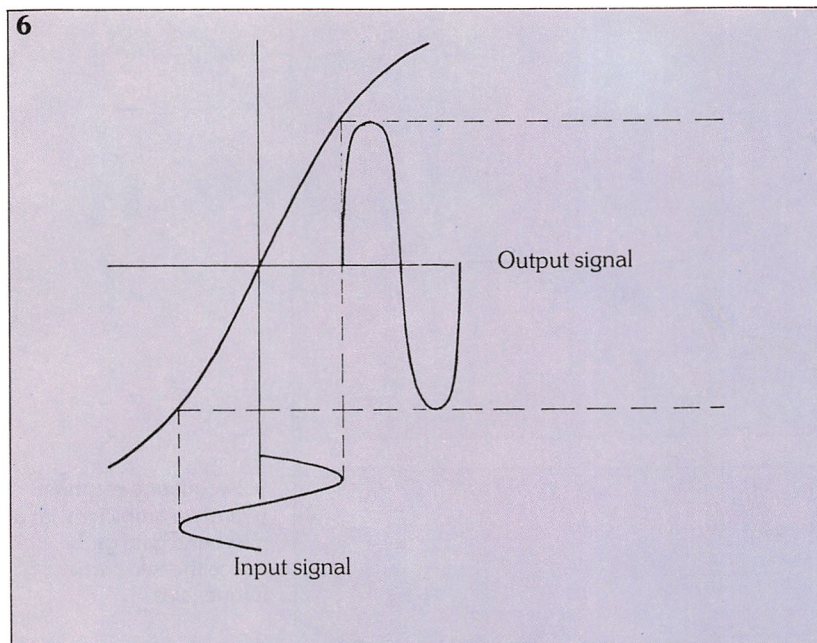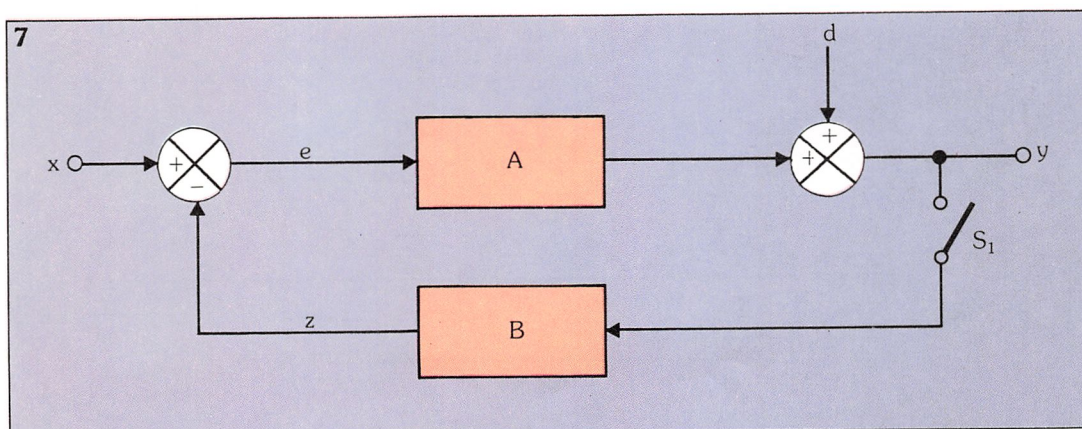*(continued from part 24)*

Distortion

Distortion in an analogue circuit may be considered to be a form of noise, where extra components, related to the signal being amplified, are generated within the circuit. One common type of distortion, known as **harmonic distortion**, results in the introduction of frequencies which are exact multiples of the fundamental frequency being amplified: these multiples are called **harmonics**.

Harmonic distortion may be introduced if the amplifying device, say, a transistor, has a curved transfer characteristic. *Figure 6* illustrates such a curved transfer characteristic together with input and output signals: the output signal is noticeably distorted due to harmonic distortion.

We can include the effects of distortion in the block diagram representation of the feedback system, as an extra voltage, d, added at a summing point as shown in *figure 7*. Without feedback, i.e. with switch S_1 open, the output of the circuit is given by:

$$y = Ax + d$$

However, with feedback, i.e. with switch S_1 closed, the output becomes:

$$y = Ae + d$$

As before:

$$z = By$$

and:

$$e = x - z$$

so:

$$y = A(x - By) + d$$
$$= Ax - ABy + d$$

Therefore:

$$y + ABy = Ax + d$$

and:

$$y(1 + AB) = Ax + d$$

so:

$$y = \frac{Ax}{1 + AB} + \frac{d}{1 + AB}$$

We can see that the distortion has been reduced by the feedback factor.

Noise

The above description of distortion, and how its effects may be reduced with feedback, may be applied to noise.

6. **Curved transfer characteristic** of a transistor, illustrating harmonic distortion.

7. **Block diagram representation of a feedback system** where the extra voltage, d, represents the effects of distortion.

However, noise is frequently introduced at an early point in the amplifier and the effects of this will be more noticeable than when it is introduced later in the system. Negative feedback may be shown to have virtually no effect on reducing the effects of noise introduced in the early stages in an amplifier.

Instability and frequency compensation
With any applied input signal frequency to an amplifier, the output signal may be shifted in phase. For example, *figure 8* shows the possible input and output signals of an op-amp circuit with a gain, G, of 2.

The input signal (shown in black) is a sine wave of frequency f, which has an amplitude of V_{in} volts, and is taken as our reference signal with a phase shift of zero. The output signal (in red), on the other hand, although of the same frequency f, is of a larger amplitude, V_{out}, and lags behind the input signal by 45°.

Now, with a gain G of 2, we might expect that the amplitude of the output signal, V_{out}, should be twice that of the
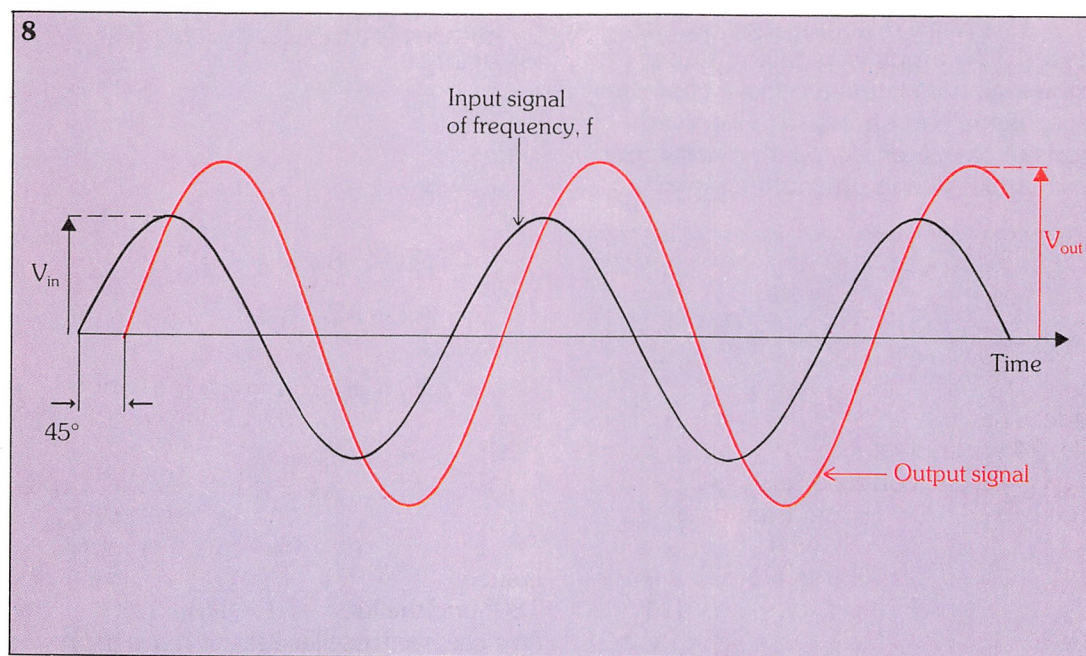
8. Input and output signals of an op-amp circuit with a gain, G, of 2. The output signal lags 45° behind the input signal.
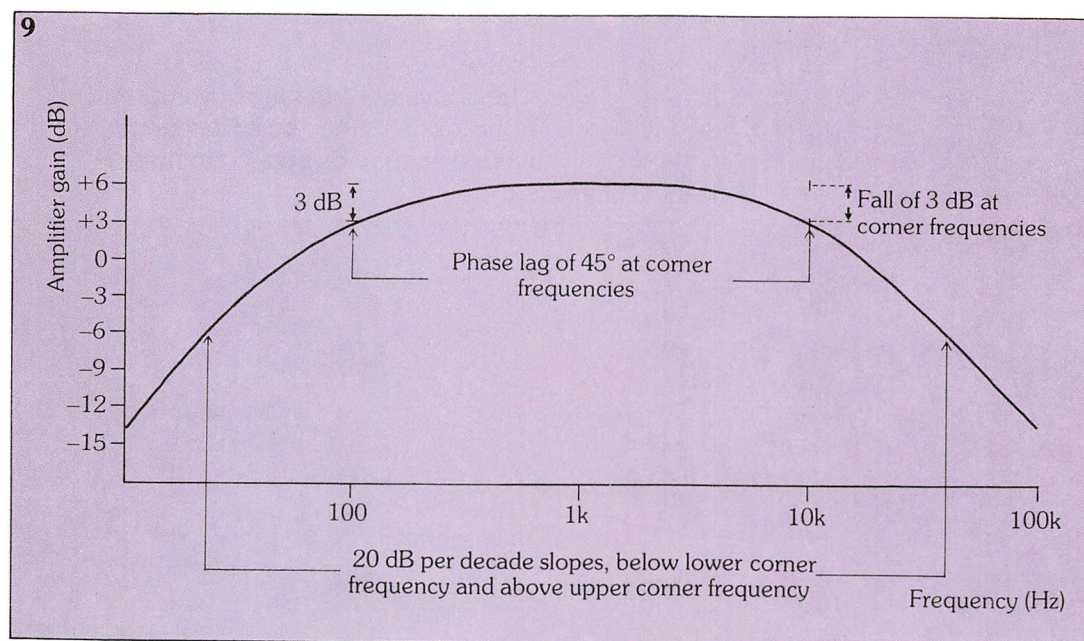
9. Frequency response plot of an amplifier with a mid-band gain of 2 – notice the two corner frequencies.

amplitude of the input signal, i.e. $2V_{in}$. However, as *figure 8* shows, this is not the case in our example because the gain we have stated is the **mid-band gain** – the gain of the amplifier in the middle of its bandwidth.

In the middle of the amplifier's bandwidth the gain *is* 2 and little or no phase shift occurs between input and output signals. But if the amplifier is used to amplify a frequency at the edge of its bandwidth, say at the bandwidth's corner frequency, then phase shift *does* occur and the actual gain is less than the mid-band gain.

If we were to measure the input and output signal amplitudes we would find that:

$$V_{out} = 1.414 \, V_{in}$$

Now, if this is related to the mid-band gain, it can be expressed as:

$$V_{out} = 0.707 \, (2 \, V_{in})$$

The '2' in the expression is the mid-band gain, G, which is multiplied by the factor 0.707 to obtain the output signal amplitude. Thinking back to *Solid State Electronics 24* we know that the factor of 0.707 is a special case (corresponding to -3 dB) which defines the corner frequency, f_c, of the bandwidth of a circuit.

Figure 9 shows a possible frequency response plot of an amplifier with a mid-band gain of 2 (which, incidentally, corresponds to $+6$ dB). Notice that the amplifier has two corner frequencies which define the bandwidth as 100 Hz to 10 kHz. Below the lower corner frequency, and above the upper corner frequency, the amplifier gain falls by 20 dB per decade, as we would expect.

Because the output signal shown in *figure 8* has an amplitude of $0.707 \times 2 \, V_{in}$, we can surmise, therefore, that the signals are at a frequency which equals one of the two corner frequencies of the amplifier's bandwidth. At both of an amplifier's corner frequencies the output signal lags the input signal by 45°, which corresponds well with our example. In the mid-band between the two corner frequencies there is little or no phase lag.

What causes instability

Although negative feedback, as we saw earlier, is an important factor in stabilising the various parameters of an amplifier, it can cause instability. How can this paradox be explained?

As we have just seen, a phase shift occurs at certain frequencies between the input and output signals of an amplifier. If the feedback network of the amplifier, or other cascaded networks, produces further phase shift, then the *total* phase shift may become 180°, i.e. inverted, and so the negative feedback becomes *positive* causing possible oscillations in the amplifier.

This can be illustrated mathematically by studying the transfer function of the amplifier:

$$G = \frac{A}{1 + AB}$$

If the whole loop, indicated by the loop gain AB, produces a phase shift of 180°, this may be shown by changing the sign of the loop gain, giving:

$$G = \frac{A}{1 - AB}$$

Now if the loop gain is unity, the denominator of the formula becomes zero and thus the gain is infinite – indicating instability.

We can therefore define the instability of an amplifier as occurring when feedback is such that the total loop phase shift is 180° *and* the loop gain is unity. The amplifier may oscillate at the frequency at which this occurs.

Why does this instability occur?

The amplifier frequency response shown in *figure 9* is produced by a stable amplifier with resistive feedback, which will not oscillate. The slope above the upper corner frequency, for example, is of a single lag response, i.e. it slopes at 20 dB per decade.

Figure 10, on the other hand, shows the upper slope of an amplifier with a more complex slope. The first part of the slope, above the first corner frequency f_1, is at 20 dB per decade – as we would expect from a single lag response. However, a second corner frequency occurs at f_2, above which the slope is at 40 dB per decade – a double lag response. Finally, a third corner frequency occurs at f_3, above which the slope is at 60 dB per decade – a triple lag
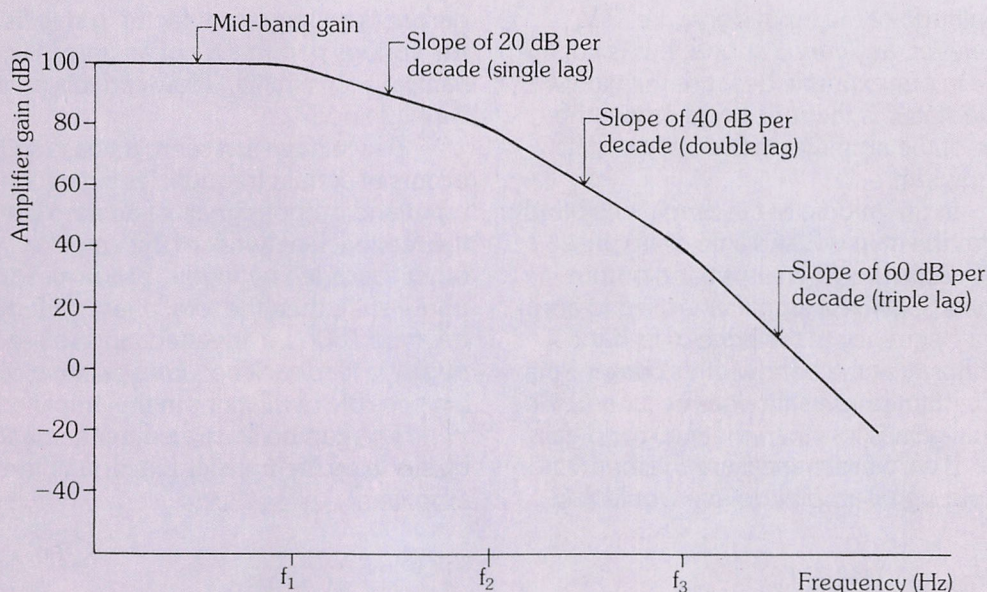
10. **Frequency response curve** of an amplifier with 3 corner frequencies.

11. Internal circuit of a 741 op-amp.

12. Open loop frequency response graph for the 741 op-amp.

response. Such a frequency response curve will occur with an op-amp which has, say, parasitic capacitances, or more than one frequency dependent component in the feedback network.

Each lag in the response curve creates a phase shift, and as each phase shift may be up to 90° it is easy to see how a stable amplifier may actually become unstable when restrictive feedback is connected, and when the total loop phase shift adds up to 180° and the loop gain is 1. Designers of amplifying circuits, therefore, need to ensure that the feedback they use to define a circuit's gain, distortion, noise, etc. does not cause such oscillation.

Preventing oscillation

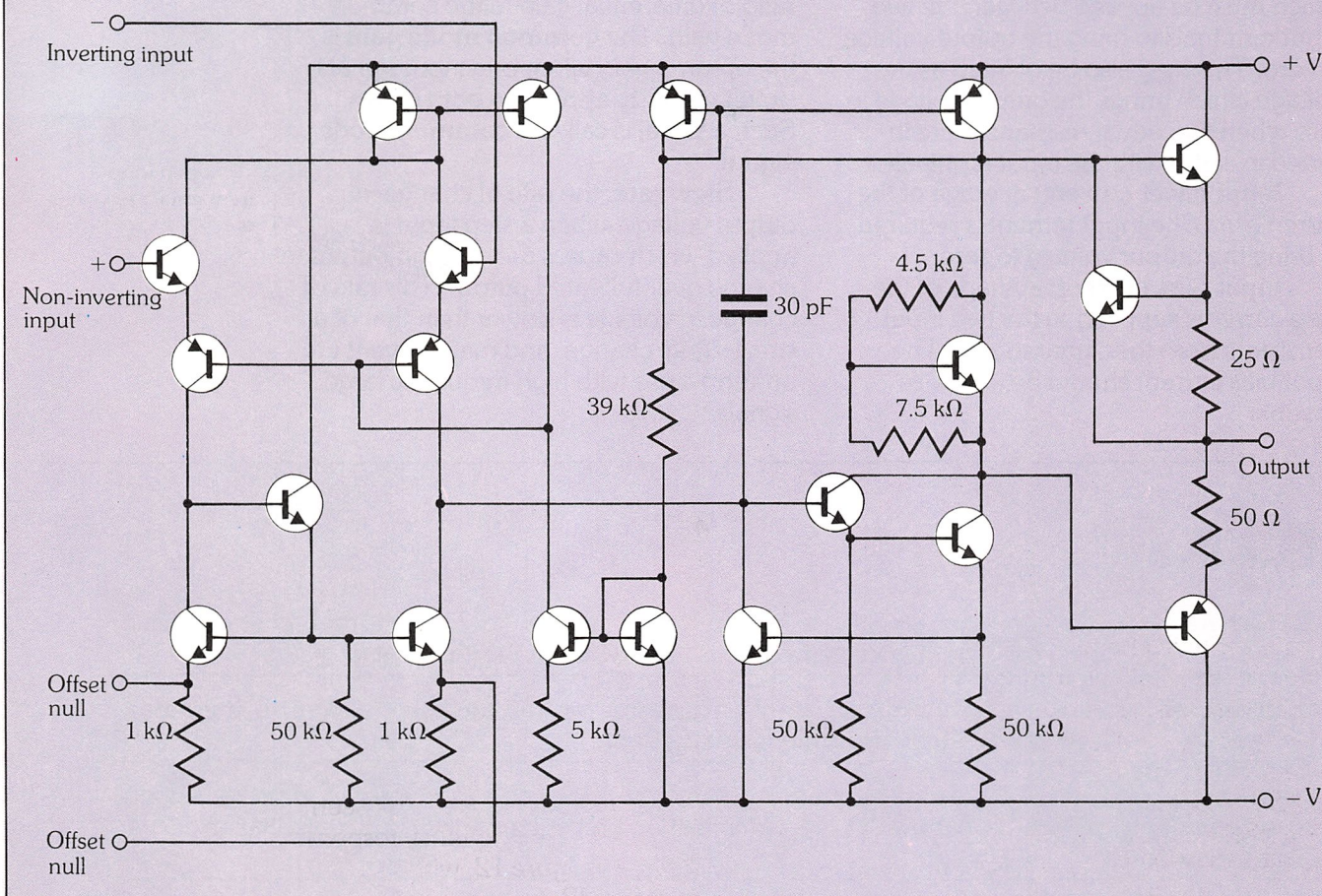The internal circuit of a 741 op-amp is shown in *figure 11* and we can see that a capacitor of value 30 pF is included in the circuit. This is a **frequency compensation** or **phase compensation** capacitor which restricts the gain of the 741 at frequencies above about 5 Hz. It is this capacitor which defines the upper corner frequency, seen in the 741's open loop frequency response graph, redrawn in *figure 12*, with its gain axis calibrated in dB.

Frequency compensation is the simplest way of reducing an amplifier's gain at those high frequencies where a loop gain of 1 may occur when the total phase shift is 180° when feedback is applied. The capacitor in the 741 circuit ensures that no instability can occur, however, it severely restricts the bandwidth at anything other than low gain, making the 741 op-amp useful only as a low frequency, low-gain (say 0 Hz to 100 kHz at 10 dB) amplifier.

Table 1
Specification for the 741 op-amp

Parameter	Minimum	Typical	Maximum	Unit
Large signal voltage gain	20,000	200,000		
Input resistance	0.2	2		$M\Omega$
Output resistance		75		Ω
Input offset voltage		1	5	mV
Input offset current		20	200	nA
Input bias current		80	500	nA
Common mode rejection ratio (CMRR)	70	90		dB
Slew rate		0.5		$V(\mu s)^{-1}$

11

12

will have a much greater bandwidth than the equivalent 741 circuit (from 0 Hz up to about 10 MHz), but the designer must add external frequency compensation capacitors, the values of which depend on the gain and desired bandwidth of the circuit.

To obtain stability with feedback, the bandwidth will be reduced to approximately the same degree as in the 741.

Op-amp specifications

The most important points in the specification of a 741 op-amp are listed in *table 1*. Each term used in the specification is defined below.

Large signal voltage gain: the ratio of the output voltage swing to the change in differential input voltage required to drive the output.

Input resistance: the resistance between the input terminals with either input grounded.

Output resistance: resistance between the output terminal and ground.

Other IC op-amps exist, with and without frequency compensation capacitors. The 709 op-amp, for example, is very similar to the 741, but without the 30 pF capacitor. This means that a 709 amplifier

Input offset voltage: the DC voltage which must be applied between the two input terminals to bring the output voltage to zero. This may also be defined as the voltage which brings the output voltage to zero when two equal resistances are inserted in series with the input terminals.

Input offset current: average of the currents into the input terminals required to bring the output voltage to zero.

Input bias current: average of the base currents supplied to the two input transistors from the signal source. The input bias current should be as low as possible.

Common mode rejection ratio: the ratio of differential gain to the common mode gain. The **common mode gain** is the op-amp gain when one input signal is simultaneously applied to both inputs. Such a signal is called a **common mode signal**.

Slew rate: the rate of change in output voltage, when a step input is applied which causes the op-amp output to change over full rated output. This rate of change in voltage is slower than that of a small signal change, and may prevent the op-amp's use with high frequency large signals.

Glossary

distortion	an amplifier's output signal is said to be distorted if it contains frequency components not present in the input signal
error signal	the error signal of an op-amp (or any feedback system) is the input signal minus the feedback signal
frequency compensation, phase compensation	use of frequency dependent feedback to reduce the effective upper bandwidth of an amplifier, to prevent instability due to oscillation
harmonics	multiples of the fundamental frequency of a signal, contained within the signal
harmonic distortion	distortion introduced by a system, due to the presence of harmonics in the output not contained in the input signal
loop gain, AB	in the gain formula of an amplifier with feedback: $$G = \frac{A}{1 + AB}$$ the quantity AB being known as the loop gain
open circuit voltage	the voltage a circuit generates across its output terminals when no load is connected to it
short circuit current	current which would be theoretically generated by a circuit if its output terminals were shorted together
slew rate	rate of change in output voltage when a step input causes the output voltage to change over full rated output range
summing point	symbol in a block diagram representation of a system which indicates the addition of signals
transfer ratio, transfer function	the mathematical expression which relates a circuit's output to its input, i.e. a circuit's gain

Resonance

When sitting in a car we often find that some panel will start to vibrate when we reach a certain speed. Increase or decrease this speed and the vibration stops. This is an example of resonance, and the vibration occurs when the natural frequency of the panel is exactly matched by the frequency of vibration of the engine. The same situation exists in electrical circuits and is used in many devices.

An ideal series resonant circuit

Figure 1a shows a circuit comprising an ideal inductance in series with an ideal capacitance. If we assume that a current of rms value I, at a frequency f, is flowing through the circuit, then we can draw the phasor diagram shown in *figure 1b*, where V_L leads and V_C lags the current by 90°. The values of V_L and V_C are:

$$V_L = 2\pi f L I$$
$$V_C = \frac{I}{2\pi f C}$$

Now, when the frequency, f, is high we can see

that V_L will be large and V_C will be small and the total voltage, V, will be as shown in *figure 1b*. Here, V leads the current through the circuit, and the circuit has an overall capacitive reactance. If, on the other hand, the frequency is low, V_L will be small and V_C will be large. Here, the total voltage will lag the current (*figure 1c*) and the circuit has an overall inductive reactance.

It is likely, then, that at an intermediate frequency we'll find that V_L and V_C will have exactly the same magnitude, and the resultant voltage will be zero, as shown in *figure 1d*. This condition is **resonance** and a circuit in which resonance is achieved is termed a **resonant circuit**. As we've seen the condition for resonance is:

$$V_L = V_C$$

substituting from the previous expressions gives:

$$2\pi f L = \frac{1}{2\pi F C}$$

The frequency, f_o, at which resonance occurs is thus given by:

$$2\pi f_o L = \frac{1}{2\pi f_o C}$$

which gives:

$$f_o = \frac{1}{2\pi \sqrt{LC}}$$

Using this argument we can see that such a circuit can, in theory, have current flowing through it at the resonant frequency even when the supply voltage is zero. However, this cannot occur in practice as all inductors possess some degree of resistance – this means that it is necessary to supply some voltage to the circuit.

A real series resonant circuit

If we take into account the resistance of the inductor then we get the circuit model shown in *figure 2a*. *Figure 2b* shows the phasor diagram for this circuit, drawn when the supply is adjusted to the resonant frequency f_o.

As expected, the total quadrature voltage is zero, and total supply voltage is given by:

$$V = V_R$$
$$= IR$$

So, voltage has to be supplied in order that current may flow in the circuit, but as long as R is small (as it should be for a good inductor) the voltage will be small. This is the resonant condition for a series circuit. As we have seen, the voltage is minimum at resonance and this indicates that the impedance of the circuit is also minimum.

1. An ideal inductance in series with an ideal capacitance.

1

Variable frequency supply

a)

I   L   C

V_L V_C

V

b)   c)   d)

V_L V_C V I V_L I V V_C V_L I V_C
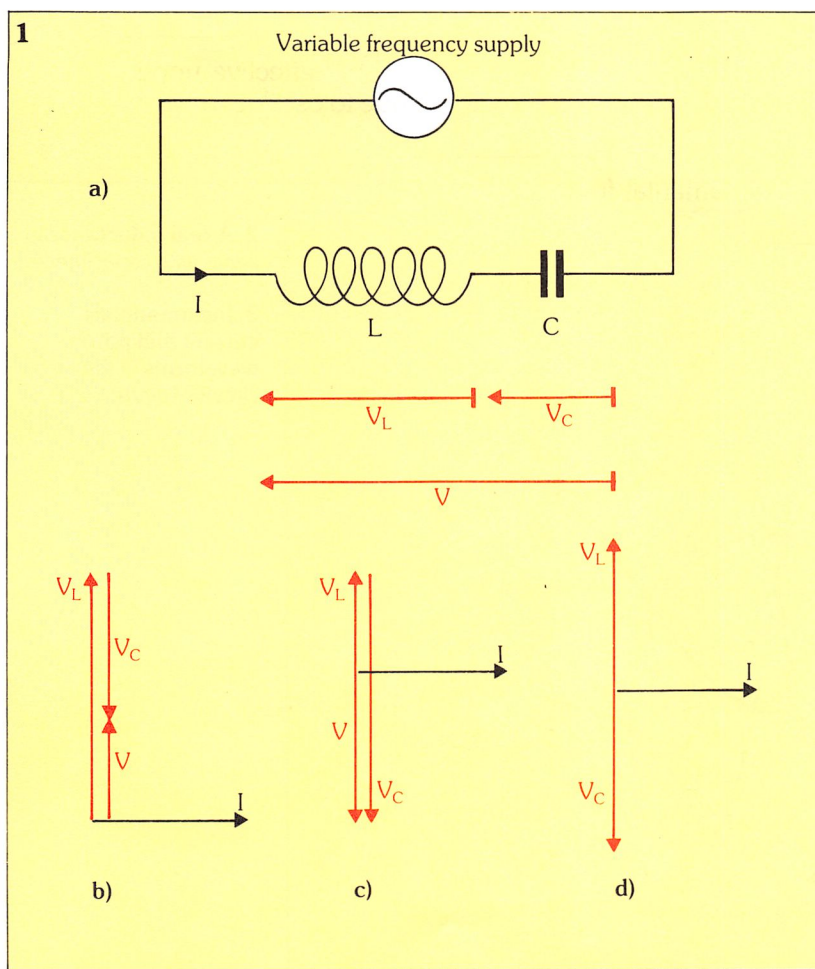
Power in a resonant circuit

Let's look at the instantaneous power flowing in the parts of this circuit. When an instantaneous current, i, flows in an inductor the power is given by:

$$p_L = v_L i$$

Similarly, the power flowing in a capacitor when an instantaneous voltage, V_C, exists across it is:

$$p_C = v_C i$$

Figure 3 shows these instantaneous power and current values as waveforms.

In the previous *Basic Theory Refresher* we saw that power flows into an inductor (positive power) when the magnitude of the current is increasing, and flows out of the inductor (negative power) when the magnitude of the current is decreasing. In the same way, power flows into a capacitor when the magnitude of the current is decreasing, and out of it when the current magnitude is increasing.

So we can see that the maximum amount of energy is stored in the inductor after time T/4, when energy has been continuously flowing into it. At this instant of time, energy begins to flow out of the inductor and into the capacitor, building up maximum energy at time T/2, when all the energy has been drained from the inductor. This continual interchange of energy is one of the characteristics of a resonant circuit.

a)

Real inductor

b)
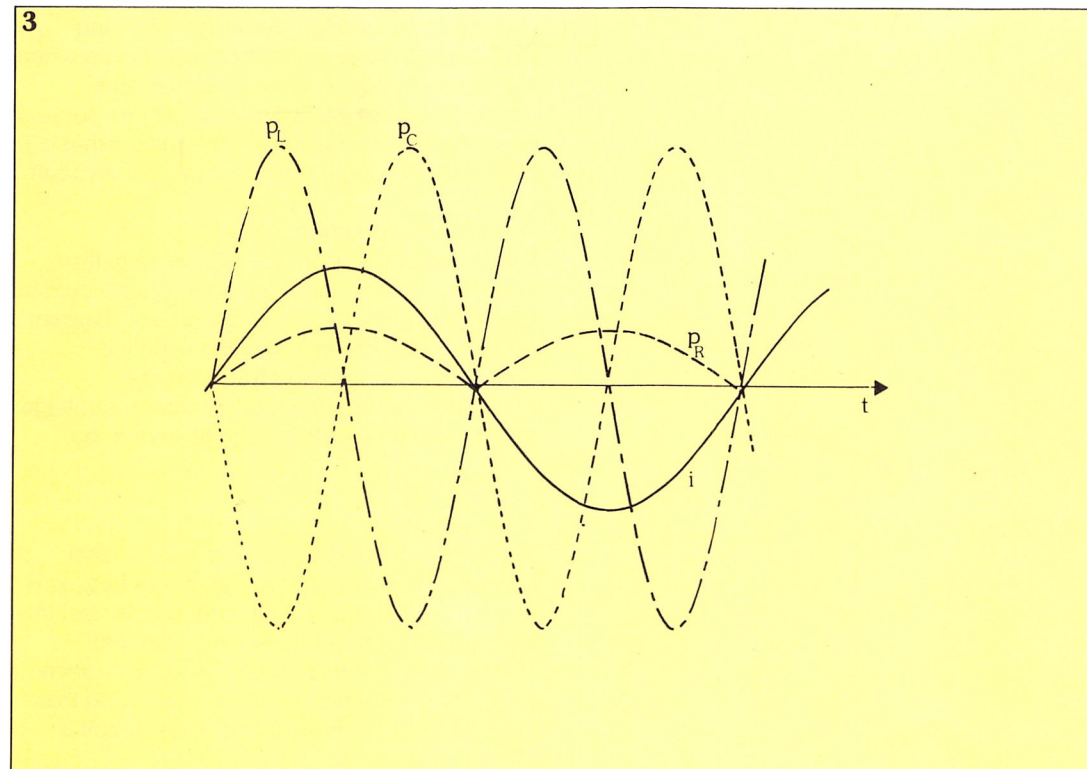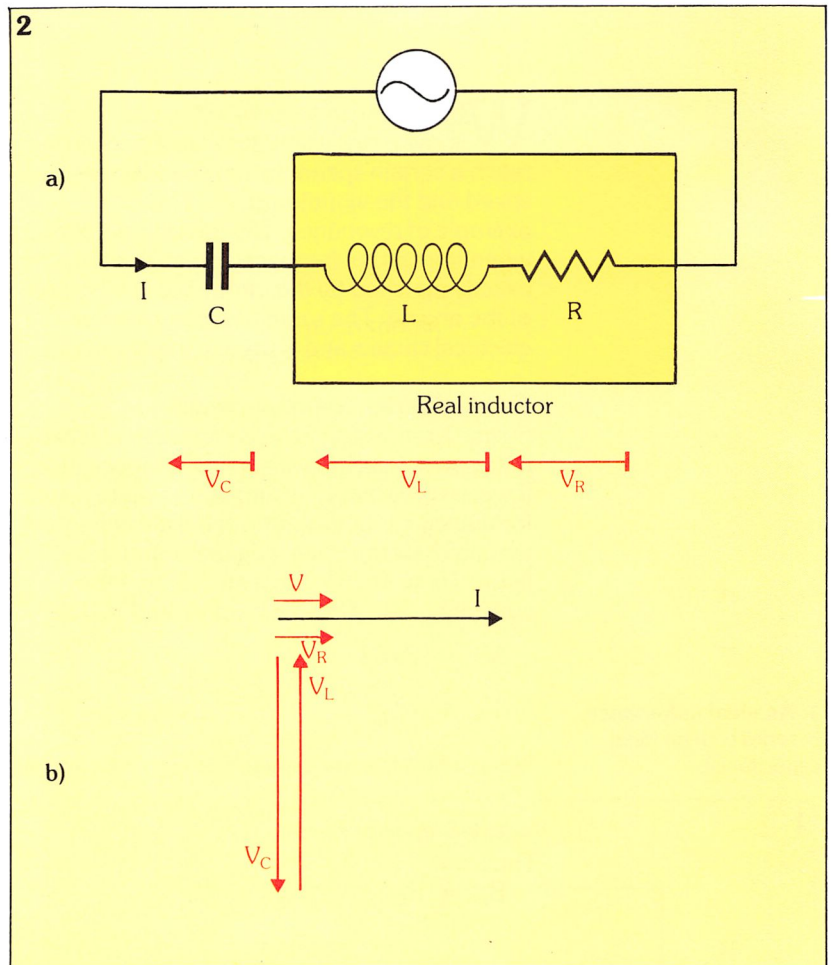
2. **A real inductance** in series with capacitance.

3. **Instantaneous current and power waveforms** in the circuit of *figure 2a*.

Considering the effect of resistance in the circuit, we know that power is dissipated in a resistance whenever current flows, and that it is always equal to:

$$p_R = v_R i$$

This is also shown in *figure 3*, and we can see that the only power which needs to be supplied from outside the circuit is that 'lost' by the resistor.

Parallel resonant circuit

Figure 4a illustrates a resonant circuit comprising an ideal inductor, capacitor and resistor connected in parallel. Drawing the phasor diagram for this circuit, we start with the voltage as reference and then build up the current through the inductor, I_L, the capacitor I_C, and the resistor, I_R, in turn. This is shown in *figure 4b* for a supply of frequency, f_o, at which:

$$I_L = I_C$$

or, in other words:

$$V_L B_L = V_C B_C$$

where B_L and B_C are the **susceptances** of the inductor and capacitor where:

$$B_L = \frac{1}{2\pi f_o L}$$
$$B_C = 2\pi f_o C$$

Thus, the resonant frequency given by:

$$B_L = B_C$$

can be obtained from:

$$f_o = \frac{1}{2\pi \sqrt{LC}}$$

From this it can be seen that the resonant frequency of a parallel circuit is the same as that for the series circuit.

If we considered a real inductor that possessed some resistance, then we would find that this expression is only approximately true, but the error is negligible if the resistance is small.

It can be seen from the phasor diagram that when the circuit is resonating the current taken from a fixed voltage supply will be at a minimum. This indicates that the impedance of the circuit will be large at resonance.

The resonant circuit and direct voltage

Figure 5a shows a series resonant circuit of a capacitor, inductor and resistor connected to a battery via a switch. When the switch is closed, a current attempts to flow in the circuit. As the current is changing, a voltage is developed across the inductor according to Lenz's law. However, since the voltage across a capacitor cannot change instantaneously, the capacitor voltage starts at zero. As the current increases, the voltage across the inductor falls and the voltage across the capacitor rises, until the current is maximum and the inductor voltage is
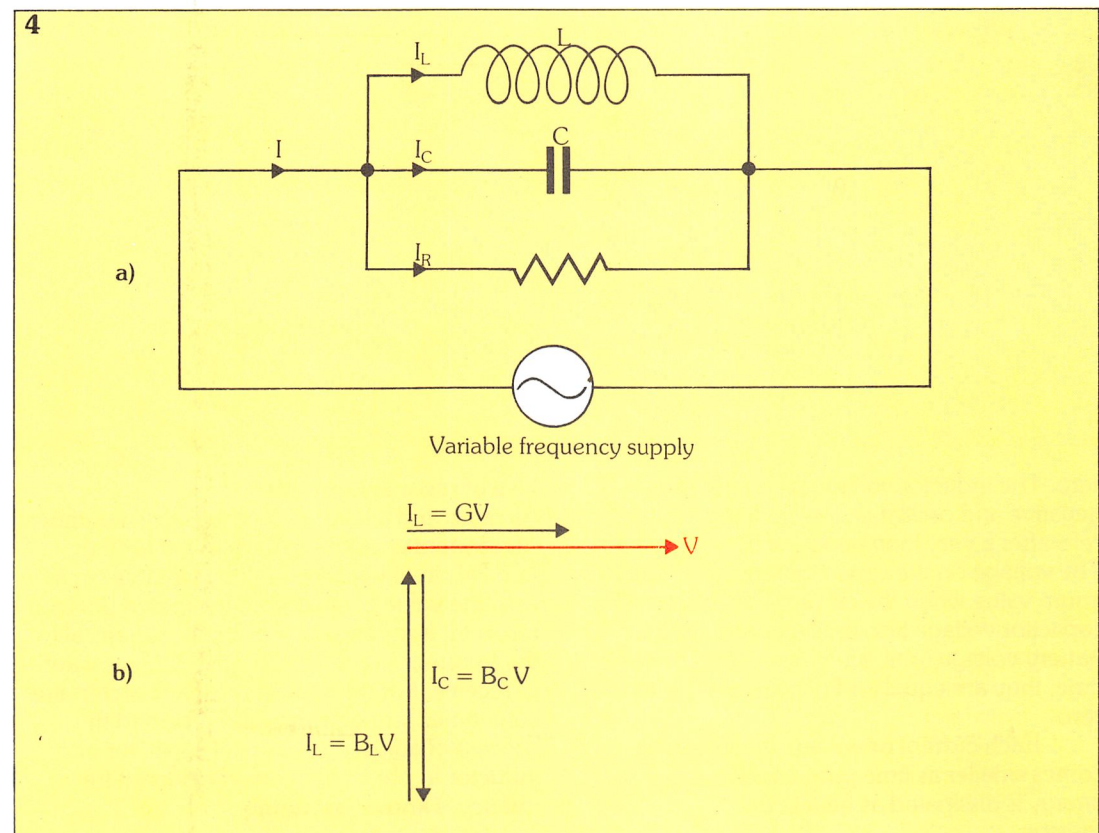
4. Circuit and phasor diagrams for a resonant circuit comprising an ideal inductor, capacitor and resistor in parallel.

5
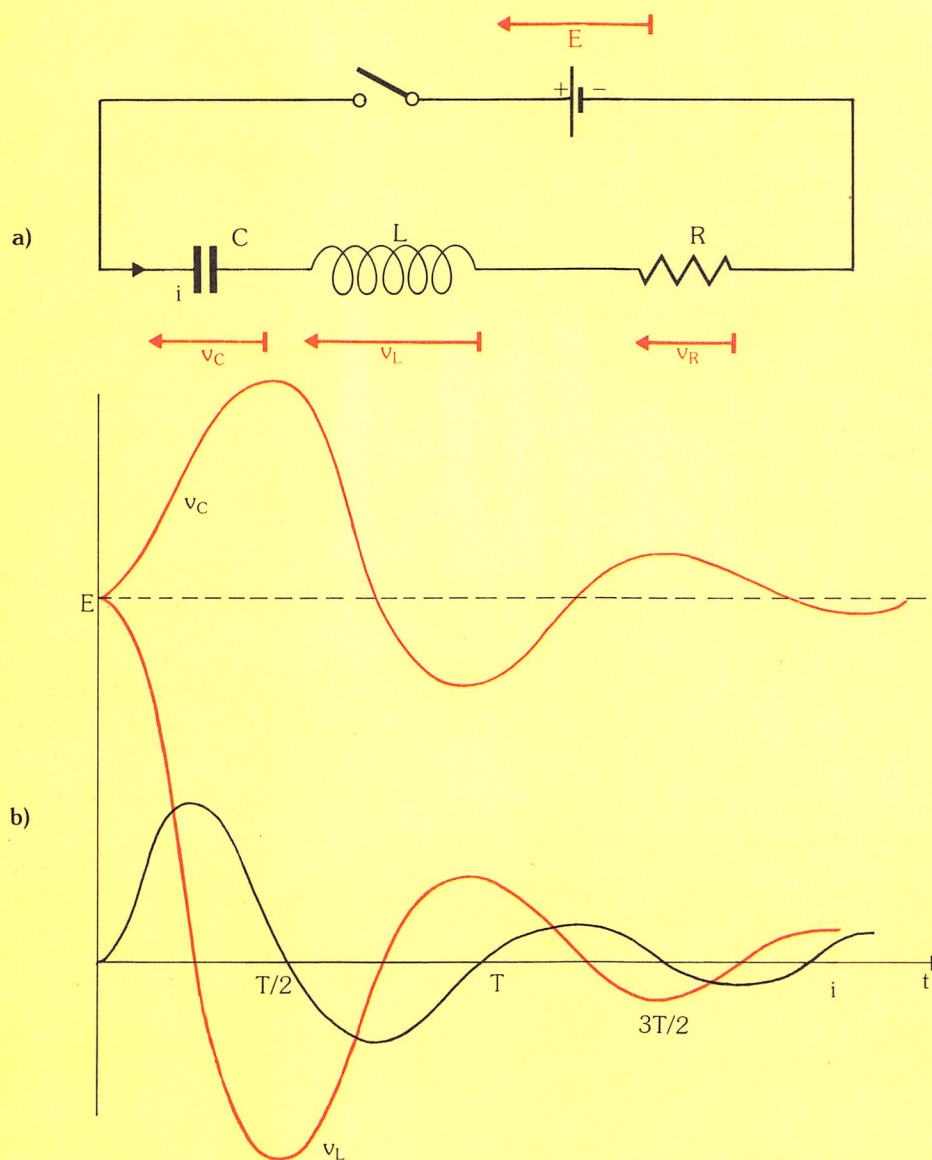
5. Circuit diagram and inductor voltage waveforms for a series resonant circuit.

a)

b)

v_C

v_L

v_R

v_C

v_L

E

i

C

L

R

E

T/2

T

3T/2

i

t

zero. The inductor voltage then becomes negative and oscillates slowly, dying away to zero after a very long period of time (*figure 5b*). The voltage on the capacitor reaches its maximum value when the current falls to zero. The capacitor voltage also oscillates around the battery voltage until, after a very long period of time, they are equal and the current falls to zero.

Each current or voltage oscillation becomes smaller as time progresses and the energy is dissipated as heat in the resistor.

Use of resonant circuits
All radio and television receivers use a number of resonant circuits – usually in the form of parallel circuits. These employ variable capacitors, the value of which can be altered until the resonant frequency of the circuit is identical to the frequency of the radio wave that we want to receive. For this reason resonant circuits are sometimes termed **tuned circuits**, and the process of adjusting either the capacitor or inductor to obtain the correct resonant frequency is known as **tuning**. □

ELECTRICAL TECHNOLOGY

Solution of general AC networks

So far, we have looked at some simple examples of the behaviour of real elements in alternating current circuits. We now need to see how more complicated networks containing a number of interconnected circuits may be solved.

The inside of a radio is a good example of how complex many circuits can be, and shows that some routine method of analysis is needed. Although we are not able to show any of the very general techniques here, we can indicate the way in which such systems can be analysed by the use of a few simple examples. This network analysis is based on Kirchhoff's two rules, which were discussed in the earlier *Basic Theory Refresher* on DC circuit analysis.

Kirchhoff's first law

Kirchhoff's **first**, or **current law** states quite simply that in any network, the sum of all currents entering every **node** (a junction between two or more wires), is equal to the sum of all currents leaving it. This is almost obvious since current is a flow of electrons and we know that electrons cannot accumulate at any point in a circuit.

This law is valid in AC circuits at every instant of time, and can be shown to hold for the phasors that represent the rms value of the currents in each branch of the network.

(Phasors will be indicated by emboldened characters and added by the rules for phasor addition.)

So we can say that at the node, P, in *figure 1a*:

$$\mathbf{I}_3 + \mathbf{I}_5 = \mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_4$$

Figure 1b shows a typical phasor diagram for this situation. The statement above would be more usually written by saying the sum of all the currents is zero, if we consider the currents entering the node to be negative and those leaving to be positive:

$$\mathbf{I}_1 + \mathbf{I}_2 - \mathbf{I}_3 + \mathbf{I}_4 - \mathbf{I}_5 = 0$$
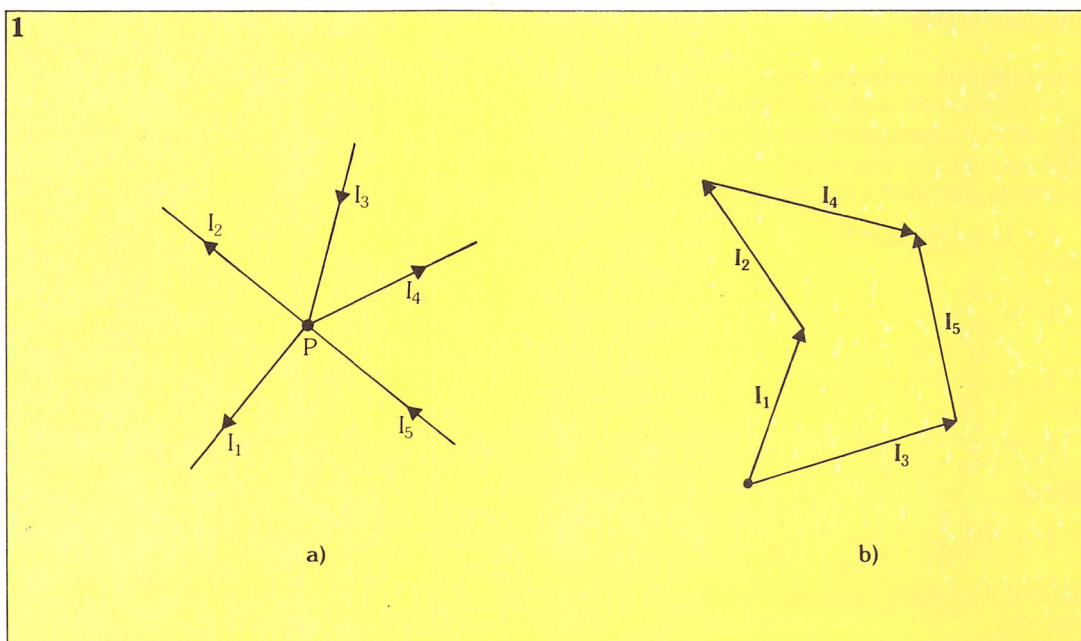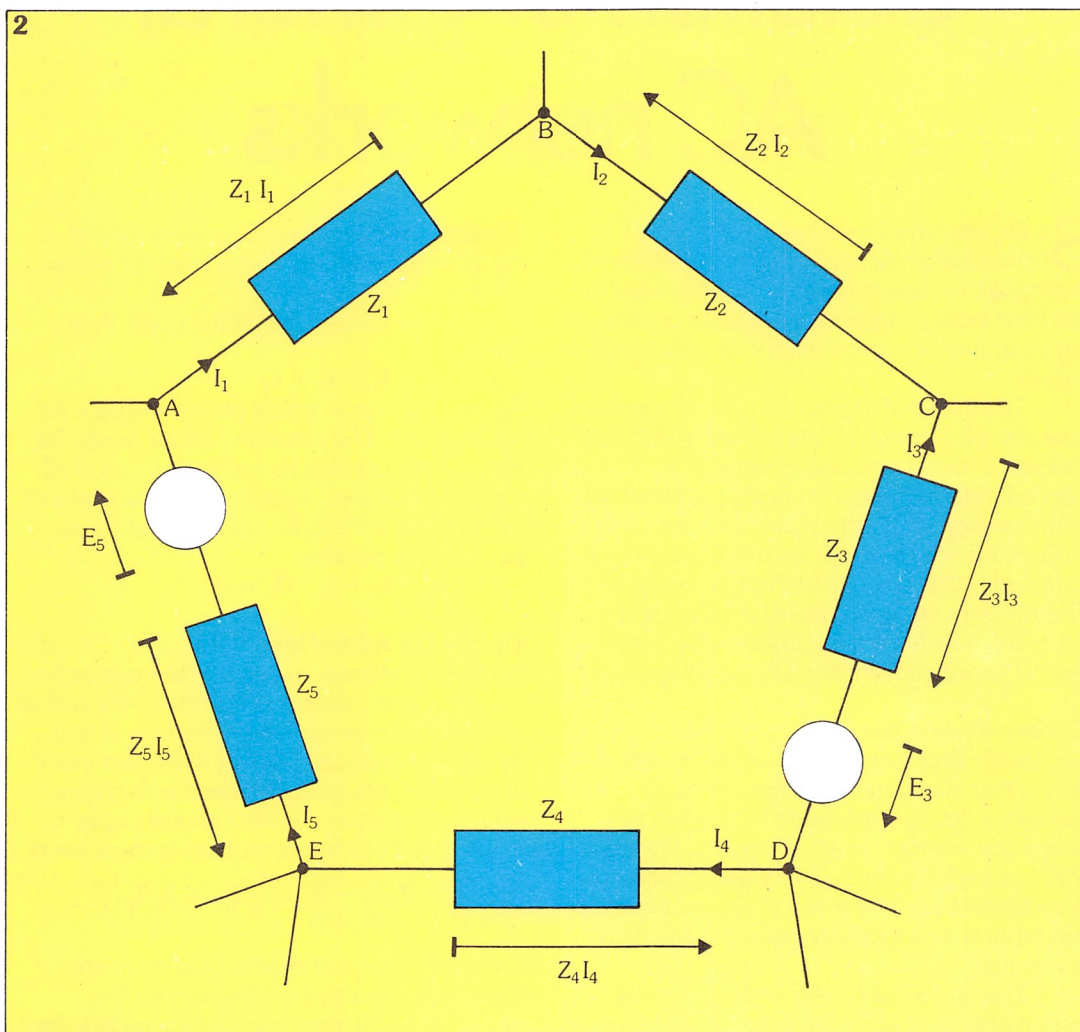
Kirchhoff's second law

Kirchhoff's **second**, or **voltage law**, states that in any network, the sum of all the EMFs and all the voltage drops across impedances in any closed loop is equal to zero. This is also true if all the currents and EMFs are phasors representing rms values. *Figure 2* shows a loop which is part of a larger network and we have:

$$\mathbf{Z}_1 \mathbf{I}_1 + \mathbf{Z}_2 \mathbf{I}_2 - \mathbf{Z}_3 \mathbf{I}_3 - \mathbf{E}_3 + \mathbf{Z}_4 \mathbf{I}_4 + \mathbf{Z}_5 \mathbf{I}_5 - \mathbf{E}_5 = 0$$

Kirchhoff's laws will enable us to solve any network, using a phasor diagram for the voltages and currents in it. Unfortunately, the solution of complicated networks by phasor diagrams is very cumbersome, but there are

1. (a) The sum of all currents entering node P is equal to the sum of all currents leaving it; **(b)** phasor diagram.

a)

b)

2

2. The sum of all the EMFs and all the voltage drops across impedances in any closed loop is equal to zero.

3. Network of three impedances connected across a sinusoidal voltage generator.

alternative methods. Let's look at two examples to illustrate the use of Kirchhoff's voltage and current laws.

Example one – current law

Figure 3 shows a network of three impedances connected in parallel across a sinusoidal voltage generator of rms value V. At node A, we can write Kirchhoff's first law in terms of:

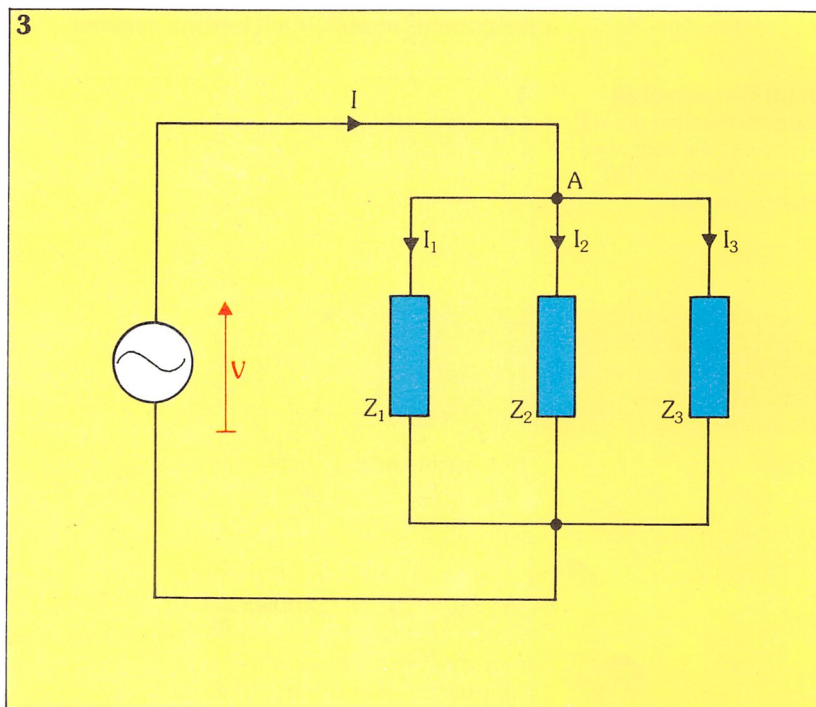$$I + I_1 + I_2 + I_3 = 0$$

since:

$$I_1 = \frac{V}{Z_1}$$

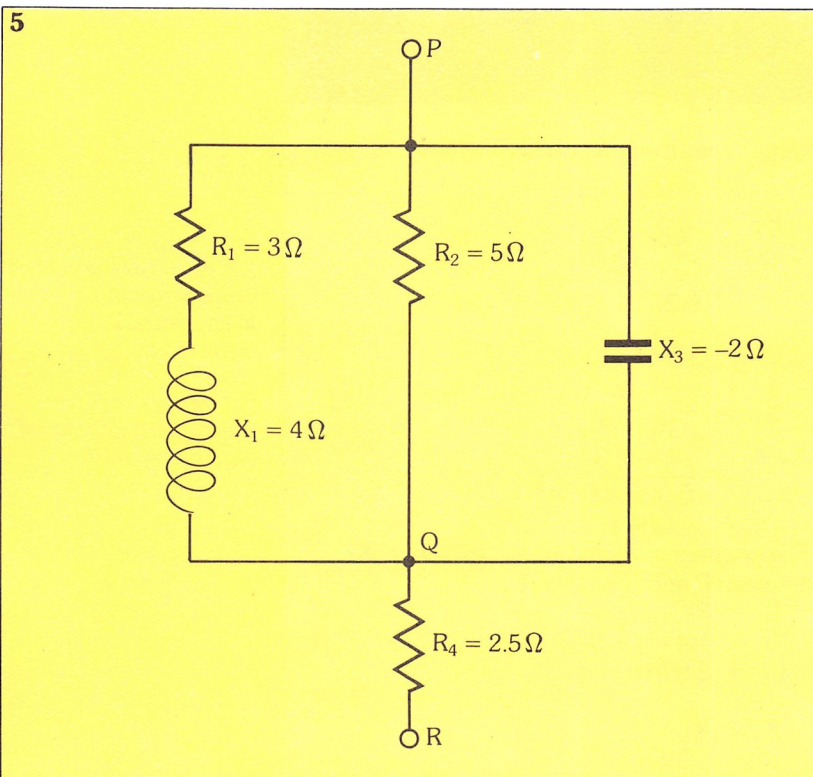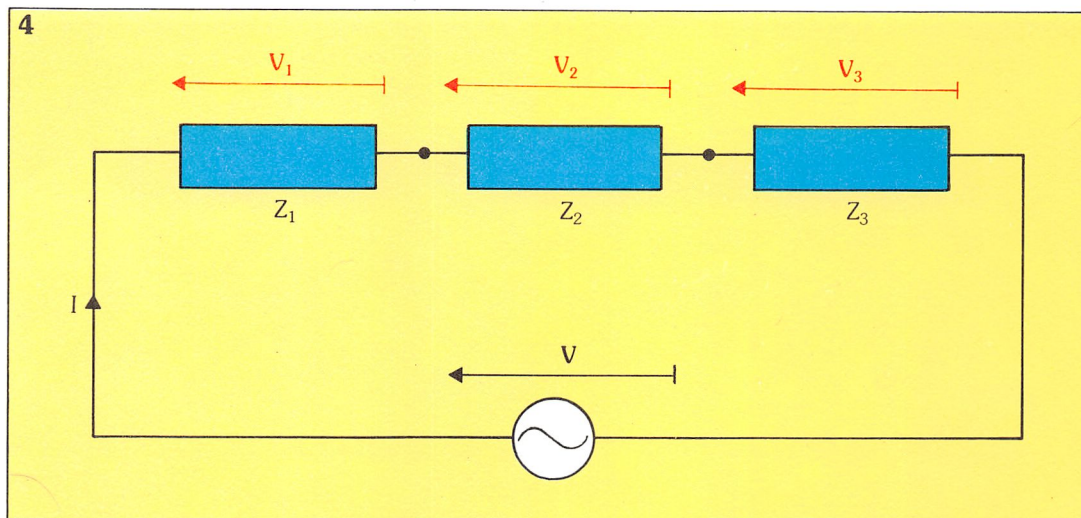$$I_2 = \frac{V}{Z_2}$$

$$I_3 = \frac{V}{Z_3}$$

giving:

$$I = V\left(\frac{1}{Z_1} + \frac{1}{Z_2} + \frac{1}{Z_3}\right)$$

Since the effective complex impedance, Z, is given by:

3

4. Single loop of three impedances connected in series across an alternating voltage.

5. Complex network of three branches in parallel, connected in series with another element.

$$Z = \frac{V}{I}$$

we have:

$$\frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_2} + \frac{1}{Z_3}$$

or by using admittances:

$$Y = \frac{1}{Z}$$

$$Y = Y_1 + Y_2 + Y_3$$

Finally, since the complex admittance is composed of in phase and quadrature parts, G and B, we can write equations for the in phase and quadrature parts of the total admittance in

these terms:

$$G = G_1 + G_2 + G_3$$
$$B = B_1 + B_2 + B_3$$

The term **complex admittance** is used to refer to the ratio of a phasor current through an admittance to the phasor voltage across it.

Example two – voltage law
Kirchhoff's voltage law can be applied to a single loop that has the same current flowing through all the impedances connected in series across an alternating voltage of rms value, **V**, as shown in *figure 4*. This gives us the phasor relationship:

$$V = V_1 + V_2 + V_3$$
$$= Z_1 I + Z_2 I + Z_3 I$$

Now, the total complex impedance, **Z**, is given by:

$$Z = \frac{V}{I}$$

Complex impedance is similar to complex admittance, and is defined as the ratio of the phasor voltage across an admittance to the phasor current flowing through it. Thus:

$$Z = Z_1 + Z_2 + Z_3$$

The complex impedance is composed of an in phase part, **R**, and a quadrature part, **X**, which may be added separately:

$$R = R_1 + R_2 + R_3$$
$$X = X_1 + X_2 + X_3$$

Solving a more complex network
Figure 5 illustrates a network that consists of three branches in parallel, connected in series with another element. We can model a circuit consisting of a resistance, R, in series with a reactance, X, by an equivalent circuit of a conductance, G, in *parallel* with a susceptance, B. and vice versa if:

$$G = \frac{R}{R^2 + X^2} \; ; \quad B = \frac{-X}{R^2 + X^2}$$

781
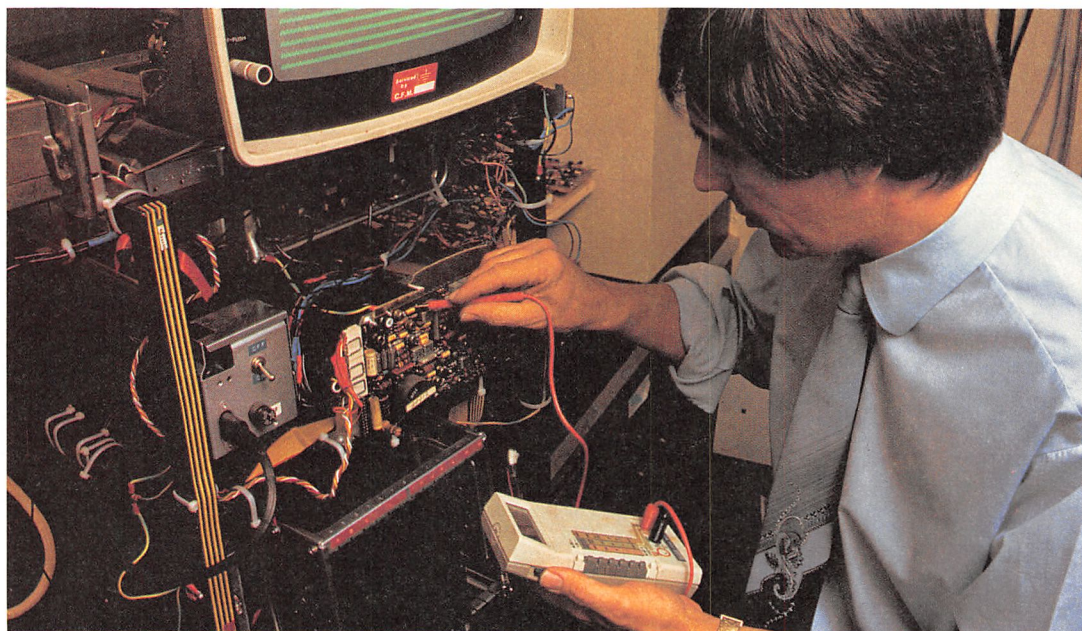
Left: testing a circuit using a multimeter. (Photo: IAL).

or, reciprocally:

$$R = \frac{G}{G^2 + B^2} \; ; \quad X = \frac{-B}{G^2 + B^2}$$

Beginning with the three branches in parallel, we can now convert these to their parallel equivalents. For branch 1: $R_1 = 3$ and $X_1 = 4$:

$$G_1 = \frac{R_1}{R_1^2 + X_1^2} = \frac{3}{3^2 + 4^2} = 0.12\,S$$

$$B = \frac{-X_1}{R_1^2 + X_1^2} = \frac{-4}{3^2 + 4^2} = -0.16\,S$$

For branch 2: $R_2 = 5$ and $X_2 = 0$, so:

$$G_2 = \frac{R_2}{R_2^2 + X_2^2}$$
$$= \frac{5}{5^2 + 0}$$
$$= 0.2\,S$$

$$B_2 = 0\,S$$

For branch 3: $R_3 = 0$ and $X_3 = -2$, so:

$$G_3 = 0$$

$$B_3 = \frac{-X_3}{R_3^2 + X_3^2}$$
$$= \frac{2}{2^2 + 0}$$
$$= 0.5\,S$$

Adding these values together gives us the total conductance, G_T, and susceptance, B_T, between the points P and Q:

$$G_T = G_1 + G_2 + G_3$$
$$= 0.12 + 0.2 + 0$$
$$= 0.32\,S$$

$$B_T = B_1 + B_2 + B_3$$
$$= -0.16 + 0 + 0.5$$
$$= 0.34\,S$$

As this is in series with R_4, it must be converted to the series form, R_T and X_T, where:

$$R_T = \frac{G_T}{G_T^2 + B_T^2}$$
$$= \frac{0.32}{0.32^2 + 0.34^2}$$
$$= 1.47\,\Omega$$

$$X_T = \frac{-B_T}{G_T^2 + B_T^2}$$
$$= \frac{-0.34}{0.32^2 + 0.34^2}$$
$$= -1.56\,\Omega$$

The resistance, R, and series reactance, X, between P and R is given by:

$$R = R_T + R_4$$
$$= 1.47 + 2.5$$
$$= 3.97\,\Omega$$

$$X = X_T + X_4$$
$$= -1.56 + 0$$
$$= -1.56\,\Omega$$

So, the impedance:

$$Z = \sqrt{R^2 + X^2}$$
$$= \sqrt{3.97^2 + 1.56^2}$$
$$= 4.27\,\Omega$$

and:

$$\cos \phi = \frac{3.97}{4.26}$$
$$= 0.931$$

and the current leads the voltage as the reactance is negative. □

Instrumentation and control systems-2

Transducers

Transducers are the devices used in electronic instrumentation and control systems to convert energy, corresponding to a measure of some physical quantity, into electrical energy, such as voltage or current.
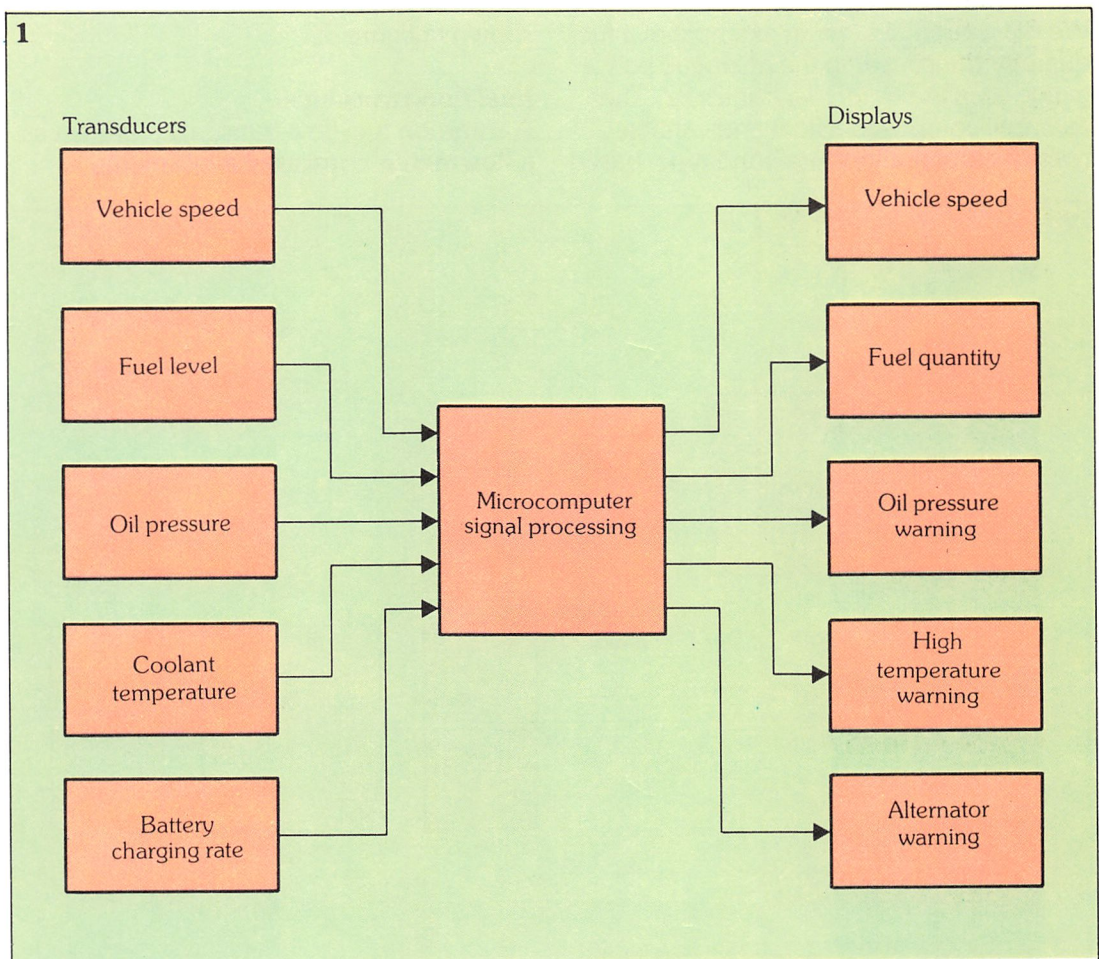
Ideally, a transducer should only be sensitive to the effects of one particular quantity, i.e. the quantity that we are using it to measure. Transducers can be manufactured to convert a sample of the energy in physical quantities such as temperature, pressure, liquid flow, distance, speed, power and height into a form which can be used by a signal processor. The signal processor, in turn, acts on the signal from the transducer to produce an electrical signal which drives the display device.

In some modern car systems, a microcomputer performs all the signal processing operations for several measurements: a possible system is shown in *figure 1*.

Generally speaking, the transducer outputs and microcomputer input are not directly compatible. Most transducers, for example, produce an analogue output signal while the microcomputer requires a digital input – an ADC is therefore required

1. Example system for a car microcomputer.

1

Transducers   Displays

Vehicle speed   Vehicle speed

Fuel level   Fuel quantity

Oil pressure   Microcomputer   Oil pressure
                      signal processing   warning

Coolant   High
temperature   temperature
                                                   warning

Battery   Alternator
charging rate   warning

for the necessary analogue-to-digital conversion.

Similarly, the digital microcomputer output is used to drive a display device of some type: if the display is a seven-segment digital display, for example, the microcomputer can drive it directly; on the other hand, if the display is analogue, an ammeter for example, then a DAC is needed to convert the digital signal to analogue form.

The microcomputer is also likely to interface between many transducers and displays, and so multiplexers of both analogue and digital varieties may be needed as shown in *figure 2*. Those physical quantities shown are only a few examples, others we may wish to measure in a fully automated car include inlet manifold pressure, crankshaft angular position, air flow rate, exhaust gas contents etc.

We will now go onto look at a few transducers in detail.

Fuel quantity transducer

We have already seen an example of a fuel quantity transducer in the analogue petrol gauge seen in *Digital Electronics 22*. That example comprised a float and variable resistor combination. A second type, based on a capacitive effect, is shown in *figure 3*, where a probe is inserted into a petrol tank. The probe is constructed of two conductive tubes mounted concentrically and separated by a known distance, d. This forms the basic capacitor arrangement of two conductive plates separated by an insulator. The capacitance of the probe is proportional to the permittivity as:
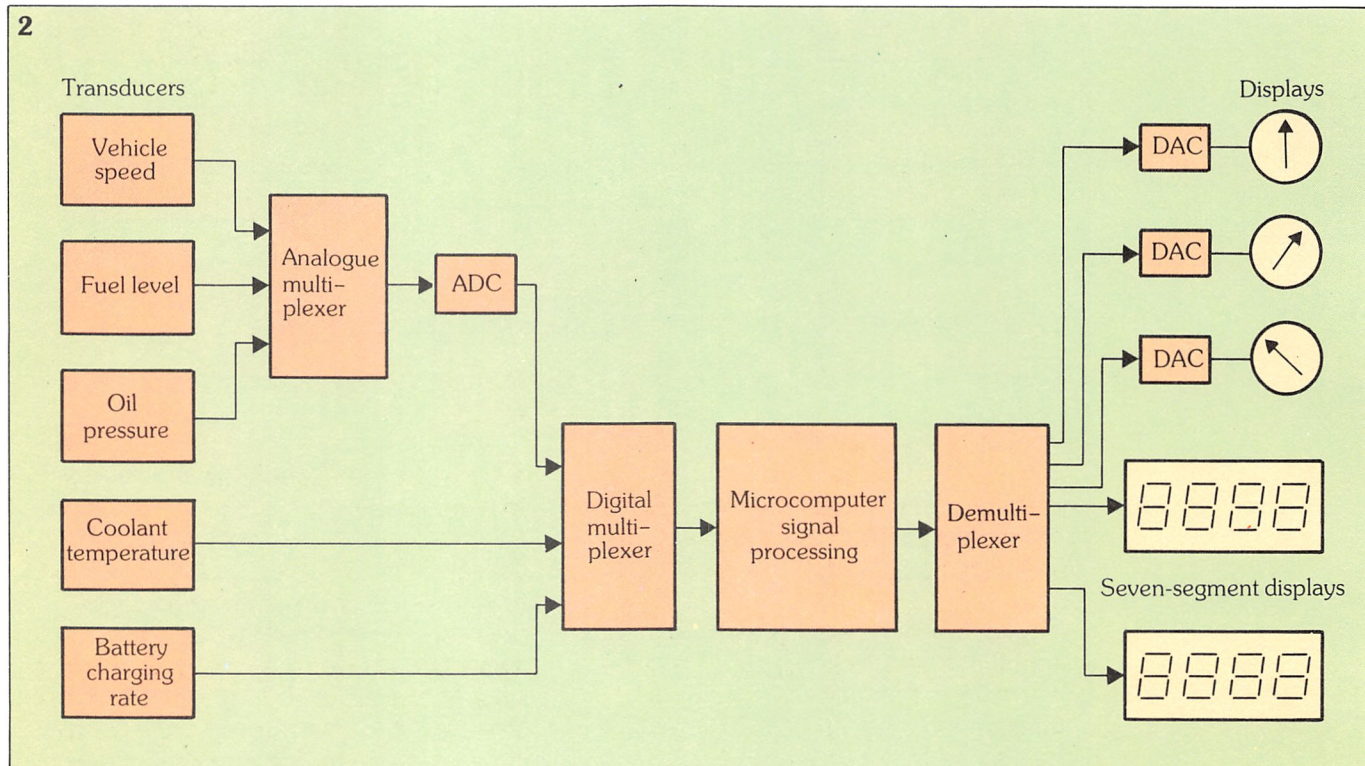
$$C \propto \varepsilon_r$$

When the tank is empty, the relative permittivity, ε_r, is the same as that of air, however, when the tank is full, the relative permittivity is that of petrol, and the capacitance of the probe changes accordingly. At petrol levels between empty and full, the capacitance of the probe lies somewhere inbetween the two extremes.

A simple way of processing the output signal of the probe transducer (i.e. its capacitance) is to use the capacitance as a frequency dependent component in an oscillator. Thus, as the level of petrol varies, so does the frequency. This is also shown in *figure 3*.
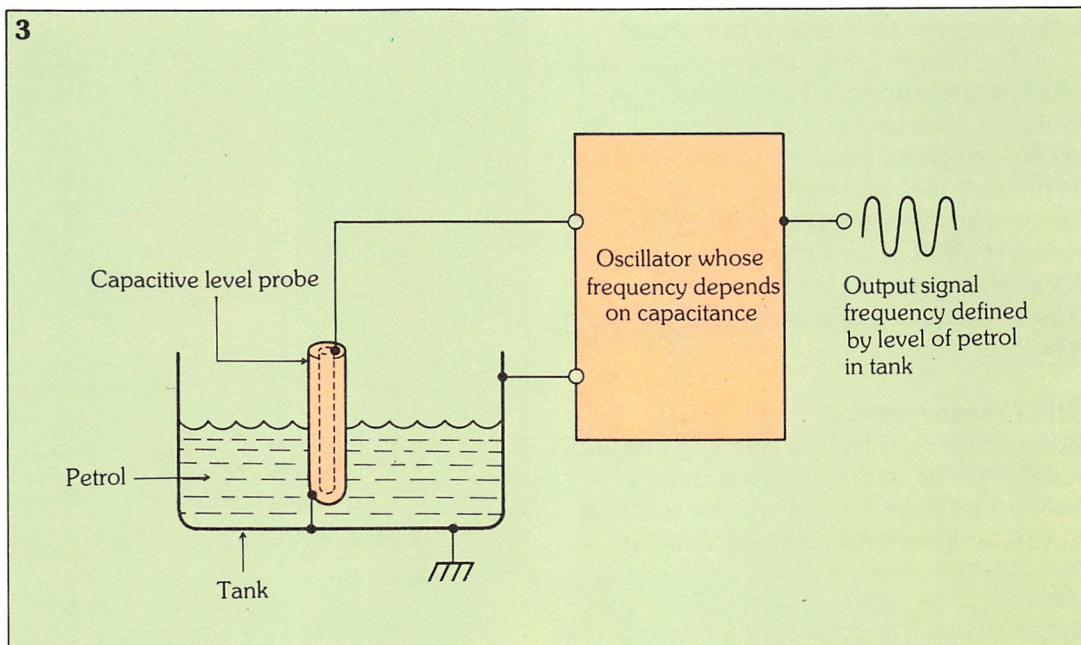
Fuel flow transducer

A common fuel flow transducer, known as a **flowmeter**, consists of a low inertia

**2. Analogue and digital multiplexers** are used to interface the car microcomputer to the transducers and displays.

2

Transducers

Displays

Vehicle speed

Fuel level

Oil pressure

Analogue multi-plexer

ADC

DAC

DAC

DAC

Coolant temperature

Battery charging rate

Digital multi-plexer

Microcomputer signal processing

Demulti-plexer

Seven-segment displays

3. Operation of a fuel quantity transducer – a capacitive level probe.

4. A flowmeter.
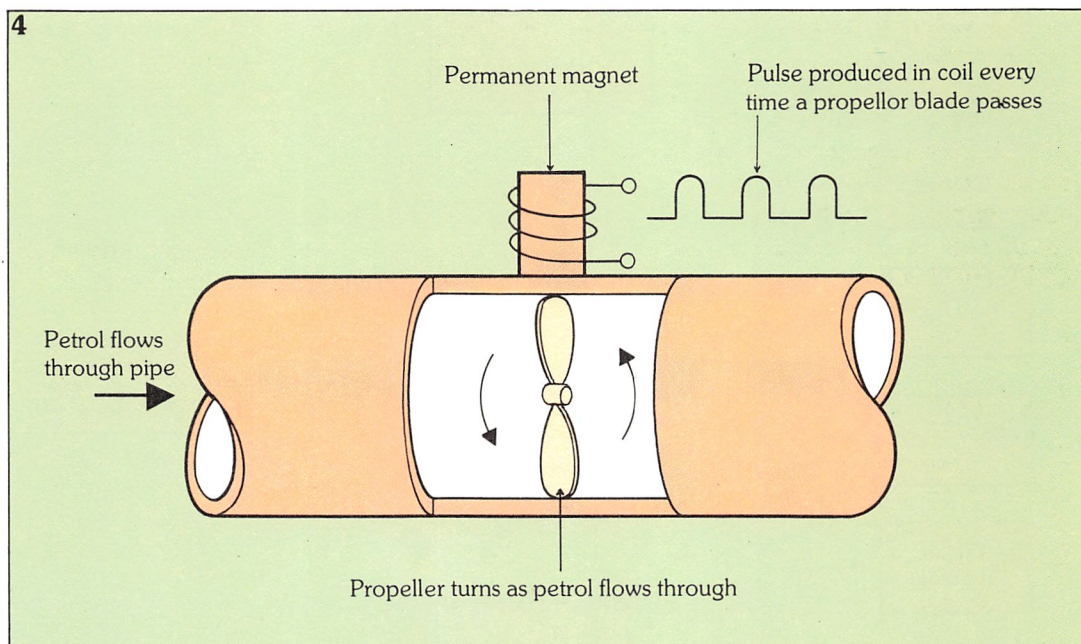
propeller, past which fuel flows as shown in *figure 4*. As the rotational speed of the propeller is determined by the fuel speed, placing the transducer in the fuel line to the carburettor enables the number of times the propeller rotates per unit time to be directly proportional to the amount of fuel used in that time.

A coil, wound around a permanent magnet, acts as a pick-up, detecting the movement of the propeller blades as it rotates. A voltage pulse is induced in the coil as each blade passes and the fre-quency of these pulses is a measure of the fuel used.
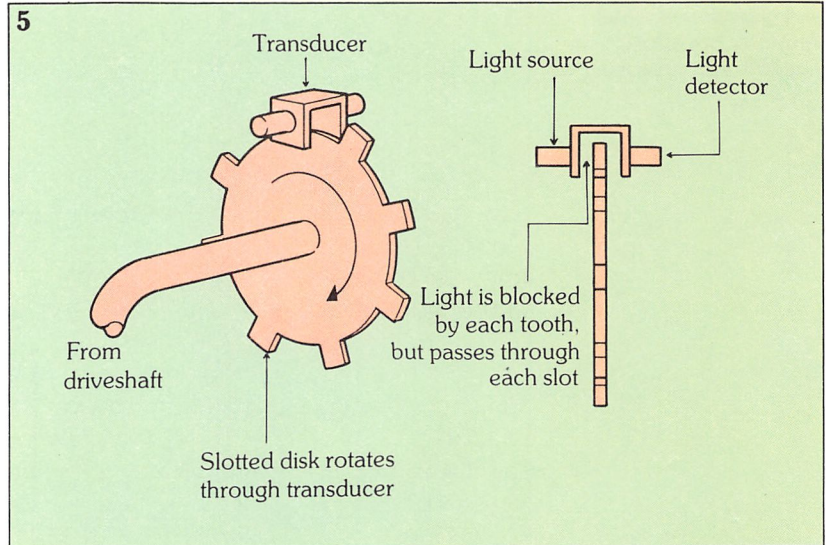
Vehicle speed transducer

In a vehicle speed measurement system, information concerning speed is mechanically coupled to the speed transducer by a flexible cable from the gearbox or driveshaft which rotates at an angular speed proportional to vehicle speed. The speed transducer generates a pulsed electrical signal for processing by the microcomputer. *Figure 5* illustrates the basic

principle behind this type of transducer.

The flexible cable drives a slotted disk which rotates through the transducer. A light source on one side of the disk shines towards a light detector on the other side. As the slots in the disk pass through the light beam, the beam is broken and the voltage output of the detector correspondingly changes. The frequency of the pulses is therefore proportional to vehicle speed.
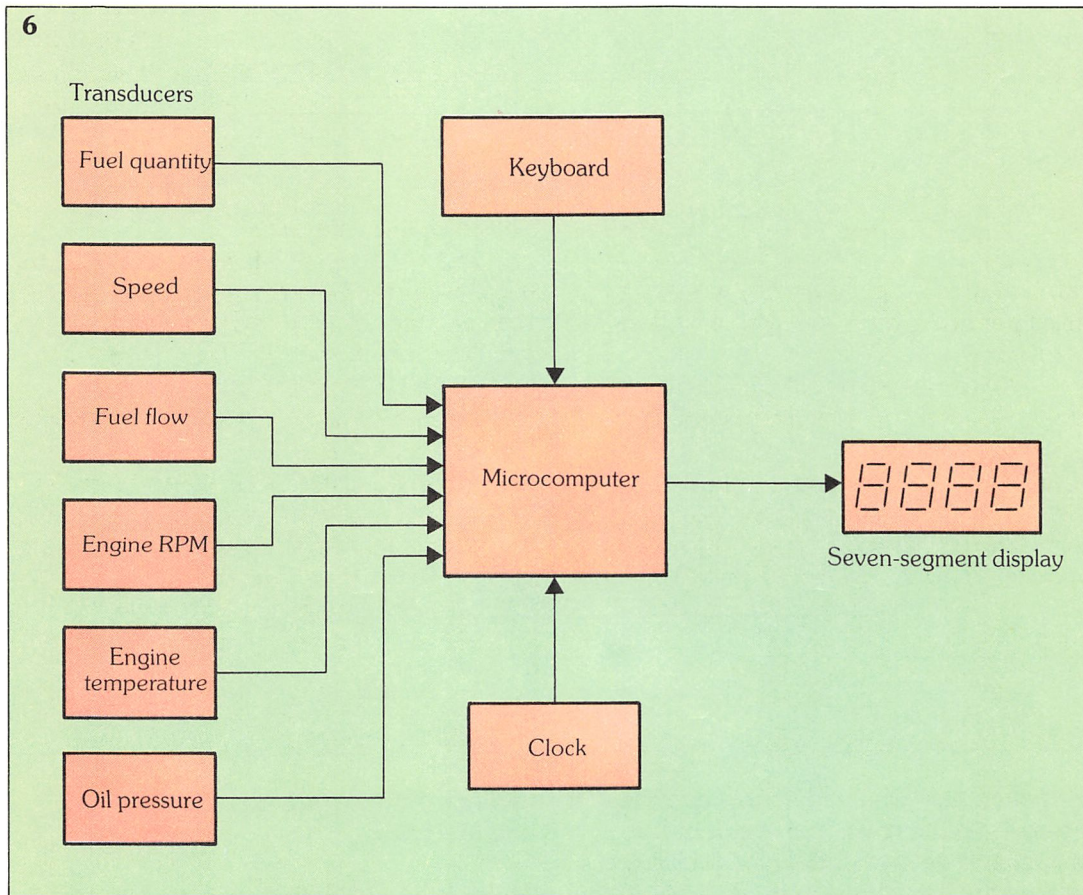
Other measurements

Measurements of battery charging rate and voltage can be undertaken without the need for transducers because the electrical analogue signals produced are directly

5. Vehicle speed transducer.

6. Possible system for a vehicle trip computer.

compatible with the ADC used.

We have seen an example of a transducer, based on the piezoresistive effect, in *Digital Electronics 22*, which can be used for measurement of oil pressure.

Engine revolutions may be measured in the same way as vehicle speed, with a flexible cable, driven by the engine crank-shaft (and *not* the vehicle driveshaft), to drive an identical transducer. Alternatively, as the number of times each spark plug is fired in a certain time equals the number of engine revolutions in that time, a count of the number of sparks in one minute equals the number of engine revolutions per minute.

A practical example

One of the most popular electronic instruments for cars is the trip information system. Such a **trip computer** may have a number of functions and can display useful information to the driver, such as: fuel and average fuel consumption; speed and average speed; distance travelled; total trip time; fuel remaining; and engine RPM, temperature and oil pressure.

All of the transducers we have discussed so far, coupled with the signal processing circuits and microcomputer, can be combined to produce a trip computer with these functions. A possible system is shown in *figure 6*.

The driver enters inputs to the system through the keyboard. At the beginning of a trip, for example, the driver initialises the system, resetting the trip distance to zero. At any time during the trip, the driver can use the keyboard to request display of particular information.

The system computes a particular trip parameter from the input data. For example, fuel consumption in miles per gallon can be found by dividing S, the speed in miles per hour, by F, fuel flow in gallons per hour.

Distance may be calculated because of the inherent relationship between distance, speed and time, given by D = S × T, where D is the distance in miles and T is the time in hours. Metric calculations, e.g. speed in kilometres per hour, can easily be performed by the microcomputer.

If each calculation is performed regularly, say, once a second, then the fuel consumption and speed calculations may be added to give running average figures, as the trip proceeds.

Control

In *Digital Electronics 22* we found that instrumentation systems are often used within control systems to provide measurements of some quantity or quantities to be controlled. The control system that we will now look at is a car speed control which can be used by the driver when travelling, say, for long distances on a motorway, to hold the car at a chosen speed. This type of control system is often known as **cruise control**.
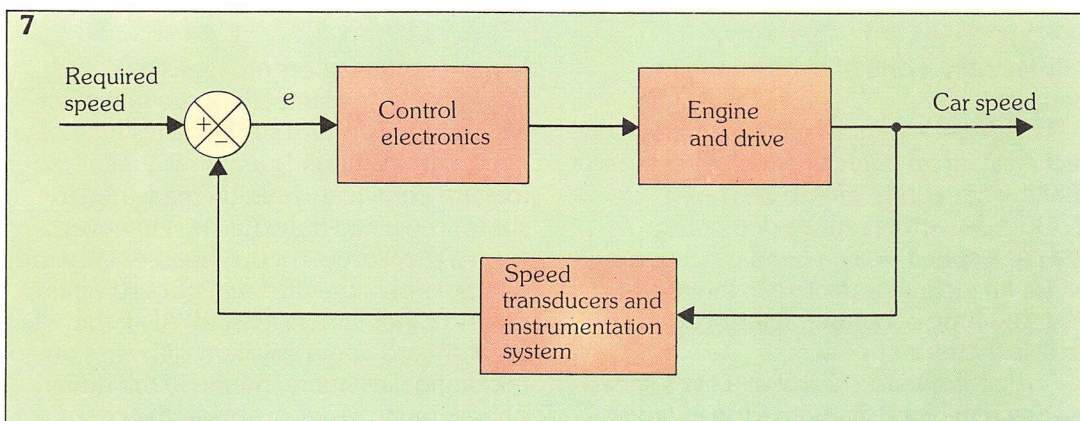
Figure 7 shows a block diagram of a cruise control system. As you can see, this is a feedback system, where a sample of the car speed is fed back via the speed transducer and trip computer instrumentation system to a summing point at the input of the system. The system's error signal, e, is thus the difference between the required speed (set by the driver) and the actual speed.
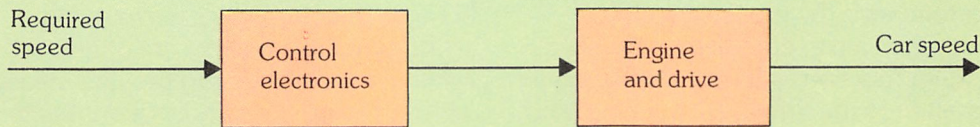
Other types of control

Feedback is not the only method of control used in control systems, **open-loop control**, illustrated in *figure 8*, is a much simpler form of control. Here, an open loop cruise control is shown, where the required car speed, i.e. how much faster or how much slower, is estimated by the driver. Let's say the driver wants to go 20 m.p.h. faster – he moves his control (say, a potentiometer knob) in the faster direction as far as he feels it is necessary to go, and the car accelerates to a higher speed.

If the driver's estimation was correct, the car will be travelling 20 m.p.h. faster than it was previously, and all appears well. However, the system does not operate
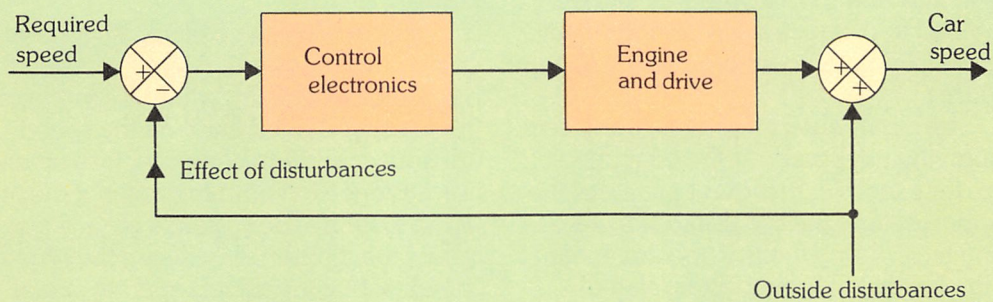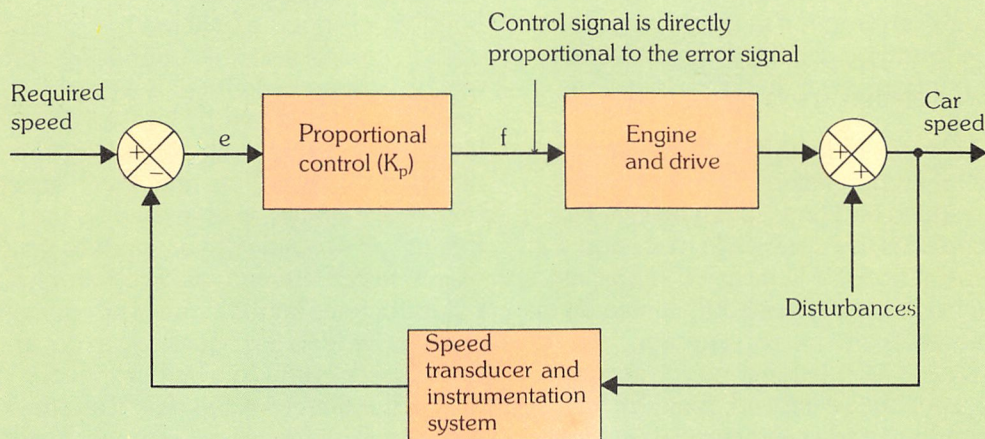
7. Block diagram for a cruise control system.

787

Required speed → Control electronics → Engine and drive → Car speed

Required speed — Control electronics — Engine and drive — Car speed

Effect of disturbances

Outside disturbances

Control signal is directly proportional to the error signal

Required speed — e — Proportional control (K_p) — f — Engine and drive — Car speed

Disturbances

Speed transducer and instrumentation system

satisfactorily if one of two main things happen:

1. The driver's original estimate was incorrect – say, if an overestimation is made so that the car accelerates to 30 m.p.h.
2. Outside factors reduce or increase the car's speed – say a head wind or a tail wind, an incline or decline in the road. Another type of control, **feedforward control**, is therefore needed.

Feedforward control of a car's speed is shown in block diagram form in *figure 9*.

Here, outside factors are measured and appropriate counteractions are taken earlier on in the control system. This type of control unfortunately assumes that *all* factors are known and can be measured in some way using transducers. However, even if measurement of a head or tail wind was possible, we still could not be certain that no other factors existed. Also, the feedforward control system still relies on the original estimate, made by the driver, of how much faster or slower the car

should be moving.

As the feedback system compares the original requirement with the car's actual speed and produces an error signal which controls the speed, it is therefore the only control system we have seen which may adequately form the basis of a car's cruise control system. In practice, feedback systems often use feedforward control too, if the system requirements demand it.

How cruise control systems work
Cruise controllers regulate car speed by adjusting the engine throttle angle, controlling the amount of fuel into the engine. Thus, the engine driving force is increased or decreased depending on whether the speed is below or above the required value. The control system must take into account the finite time lag between the newly required speed input signal and resulting final speed output signal. If the control system tries to correct speed errors too quickly, the vehicle speed may become unstable, and oscillate above and below the required value. In control terms this oscillation is known as **hunting**. If the control system corrects speed too slowly,

on the other hand, speed change will not be fast enough for practical use.
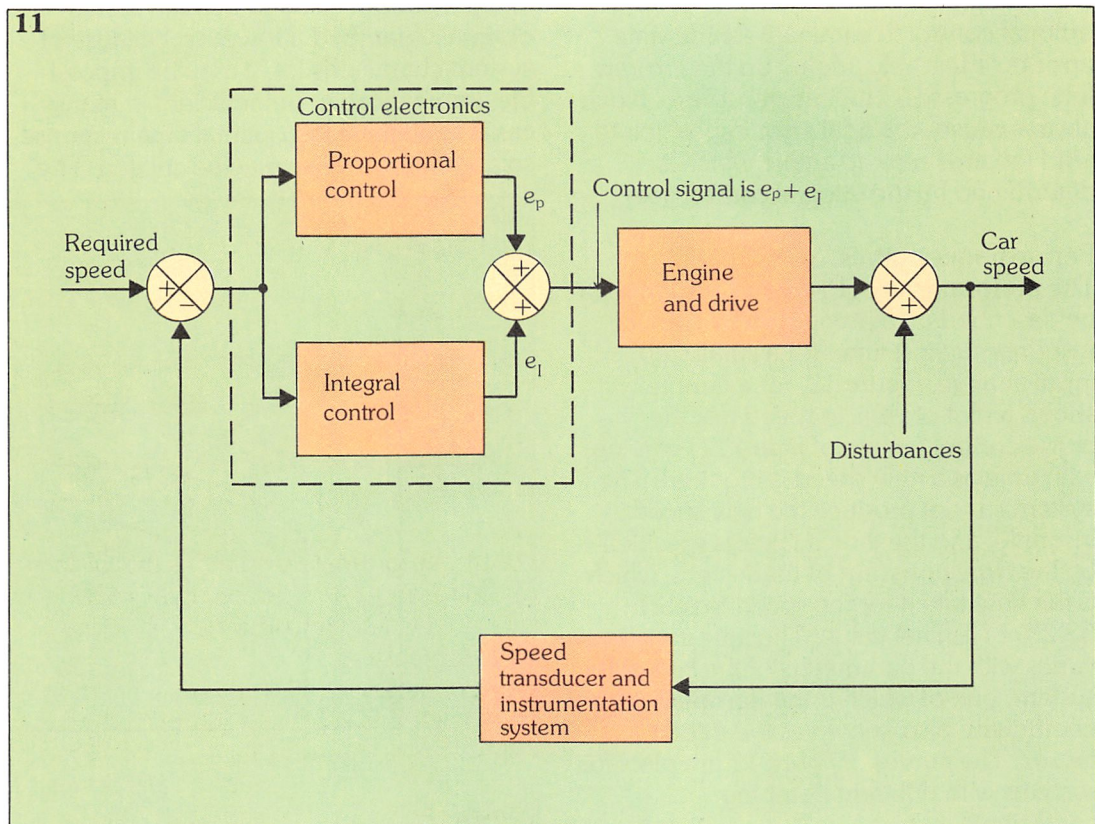
Proportional control
Feedback control systems, sometimes known as **closed loop control** systems, may be grouped into a number of different categories. The categories are generally named by the type of circuits used in the control electronics block of the block diagram. The simplest type of closed loop control system is the **proportional control** system. In this system, shown in *figure 10*, the control electronics block produces an output which is directly proportional to its input. So, the control signal is related to the error signal by:

$$f = K_pe$$

where K_p is the **proportional gain** of the control system.

Although proportional closed loop control is not complex and proportional gain may be changed quite readily to suit the application, it does have one major drawback in the **steady state error** which occurs when the system comes to rest. Steady state error is a result of the inertia of

11. Proportional – integral control system.

the system to produce a control output even if an error exists at the input. Often, steady state error is small and may, perhaps, be neglected. It depends on the proportional gain constant, K_p, and by increasing the value of the constant the steady state error may be reduced. However, increases in proportional gain may cause instability and hunting.

A much better method of eliminating steady state error is to use another category of feedback control known as **integral control**, along with proportional control. Such control systems are often termed **proportional-integral control** systems, or **PI control** systems.

Figure 11 shows a PI control system for a cruise control, in which the overall control signal is actually the sum of a proportional control signal, e_p, and an integral control signal, e_I. (The integral control signal is equal to the sum of all the error signals which have occurred since the start of the control process, taking account of their signs.) The gains K_p and K_I are chosen so that the system has a quick response, is accurate, and has no instability or oscillations.

Integral control is used with proportional control to eliminate steady state error by effectively adding up the error as time progresses. The integral control block always causes the final error in position to tend towards zero, in a time which is determined by the integral control gain.

Performance curves
The performance of PI control systems can be described by plotting system output response against time, for a small step input change. *Figure 12*, for example, shows a plot of vehicle speed (for the cruise control system of *figure 11*) varying with time as a new speed is required. The system cannot produce the new speed instantly, and the time it takes is specified by the **time constant** of the system, which is the time taken for the speed to reach 63.2% of its final value. The time constant varies with the parameters of the complete system, one of which is the **damping coefficient**, also known as the **decay factor**. The curves of *figure 12* are plots for systems with different damping coefficients.

Curve 1 represents the speed response of a system that is **over damped** – the speed of the system rises sluggishly and takes a long time to reach the required speed. It has a time constant determined by point A. Curve 4 represents the speed response of a system that is **under damped**. The speed curve rises quickly, but overshoots the required speed, and then oscillates around the speed until it finally settles down. The frequency of these oscillations is known as the system's **natural frequency**. The time constant is determined by point D.

Curves 2 and 3 represent systems that have damping coefficients that make the system **critically damped** or **optimally damped**. The critically damped system rises smoothly to the required speed with no overshoot and in a minimum amount of time; the optimally damped system rises quickly and overshoots a little, but settles quickly to the required speed. This system is called optimally damped because the vehicle speed follows the required speed more closely than for any other system.

Usually a control system designer attempts to juggle the proportional and integral control gains so that the system is optimally damped. However, because of system characteristics, it may be impossible, impracticable, or inefficient in many cases to achieve the optimal time response, so another response may be chosen. The

12. A plot of vehicle speed against time for 4 systems with different damping coefficients.

790
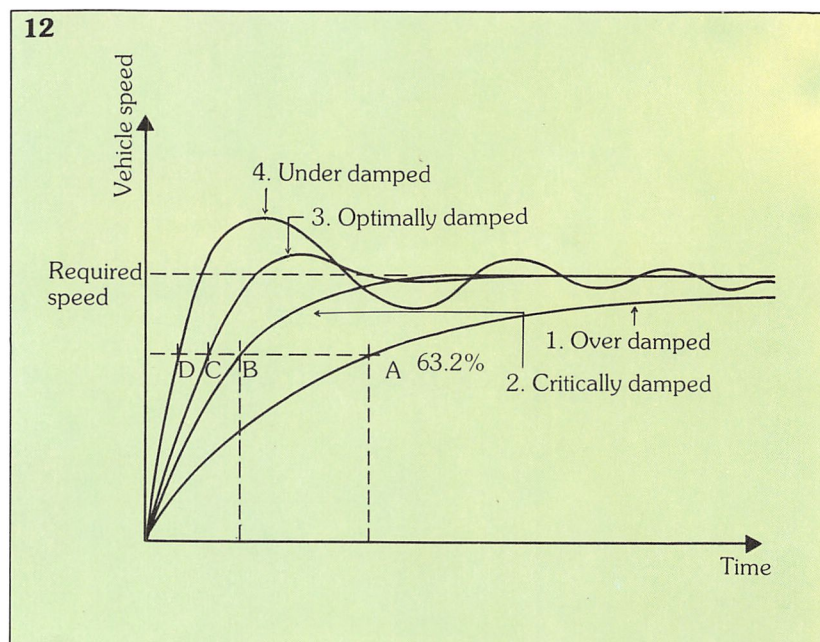
control system should make the engine drive force react quickly and accurately to the required speed, but should'nt use excessive effort (overtax the engine) in the process. The system designer, therefore, must choose the control electronics based on the following system qualities:
1. quick response;
2. relative stability;
3. small steady state error;
4. optimization of control effort required.

Frequency response curves
Another method of describing system performance is to show the response of the system to command signals, over a range of frequencies. Two types of curve, gain response and phase response, are shown in *figure 13* and these are typical of such control systems. *Figure 13a* shows gain response curves of a system, with overall or **closed loop system gain** K_s, for different damping coefficients.
The use of these curves, strictly

speaking, assumes that any changes in input value are expressed in sine wave fashion. This may seem a restrictive condition, but, in fact, it still allows the definition of any control system's performance with reasonable accuracy even if sine wave input changes are not the norm.

When the input signal frequency is low, the amplitude of the system gain is 1 and the operating point is at point X on the curve. Frequency on the curves is expressed in terms of angular frequency ω, of the signal, and is related to the actual frequency f, by the formula:

$$\omega = 2\pi f$$

As the input frequency is increased, the system gain remains constant for a certain band of frequencies and then starts changing. Note that one system with different damping coefficients has different frequency response curves. For underdamped systems, the system gain actually increases over the low frequency gain at point X. This increase corresponds to the system response shown in curve 4 of *figure 12*. When the speed command is abruptly changed, the system overshoots and oscillates around the required speed setting. A similar correspondence between the other curves of *figures 12* and *13* exists. The sluggish response of an overdamped system, indicated by curve 1 of *figure 12*, corresponds to the overdamped system frequency response of *figure 13a*, where the system gain is not as good at higher frequencies as even the critically damped system.

This is further substantiated by the curves of *figure 13b* – the phase response curves. These curves tell us how well the system output is following the input. Phase can be thought of as the number of degrees the output shaft lags the input shaft in one revolution.

The closed loop gain, K_s, of the system depends on the proportional gain, K_p, and integral gain, K_I. In order that the electronic control circuits may be properly designed, K_p and K_I must be known. The integral gain constant, K_I, depends on the mass of the engine being controlled and the natural frequency of the system.

The proportional gain depends on

13. (a) Gain response; and (b) phase response curves for a system with closed loop system gain, K_s, for different damping coefficients.
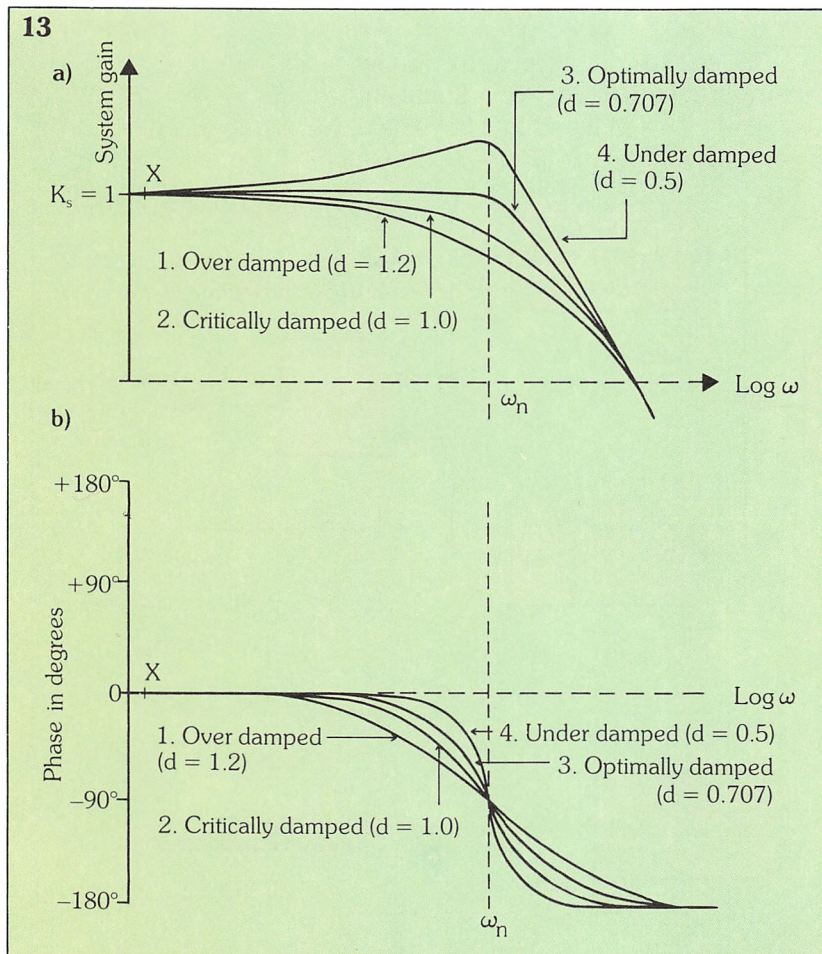
the above factors and also on the damping coefficient and the frictional force due to road resistance, wind resistance etc.

Design choices

For any automotive electronic speed control, the system designer makes several choices. The decision process might follow this pattern: first, the mass of the vehicle, the engine power, and the friction factor, are determined; second, the natural frequency, ω_n of the system is chosen. The input signal frequency range to which the controller must respond is based on a study of how quickly a human driver can detect and correct for vehicle speed changes on a hilly road surface. It is assumed that the cruise control system must respond at least as fast as the driver. Humans respond in tens of milliseconds to tenths of seconds. From ω_n, K_I is set knowing the mass of the vehicle.

Third, the damping coefficient for the kind of response desired is chosen from the values as given in *figures 13a* and *13b* using *figure 12* as a guide. Knowing these parameters, K_p is determined.

A complete cruise control system

Figure 14 shows a block diagram of the control electronics block of a possible car cruise control system. The input speed command, i.e. the required speed, and the actual car speed are stored directly in counters.

The actual speed signal is sampled periodically and is formed by the output from the speed instrumentation system discussed earlier. Required speed is set by the driver. Both numbers are compared (actually subtracted from each other to obtain a difference) and a number representing the error is applied to two paths.

In one path, a look-up table is used to produce an output number representing the system's proportional gain output, e_p. The other path digitally integrates the error

14. **Block diagram** of the control electronics block for a cruise control system.

792
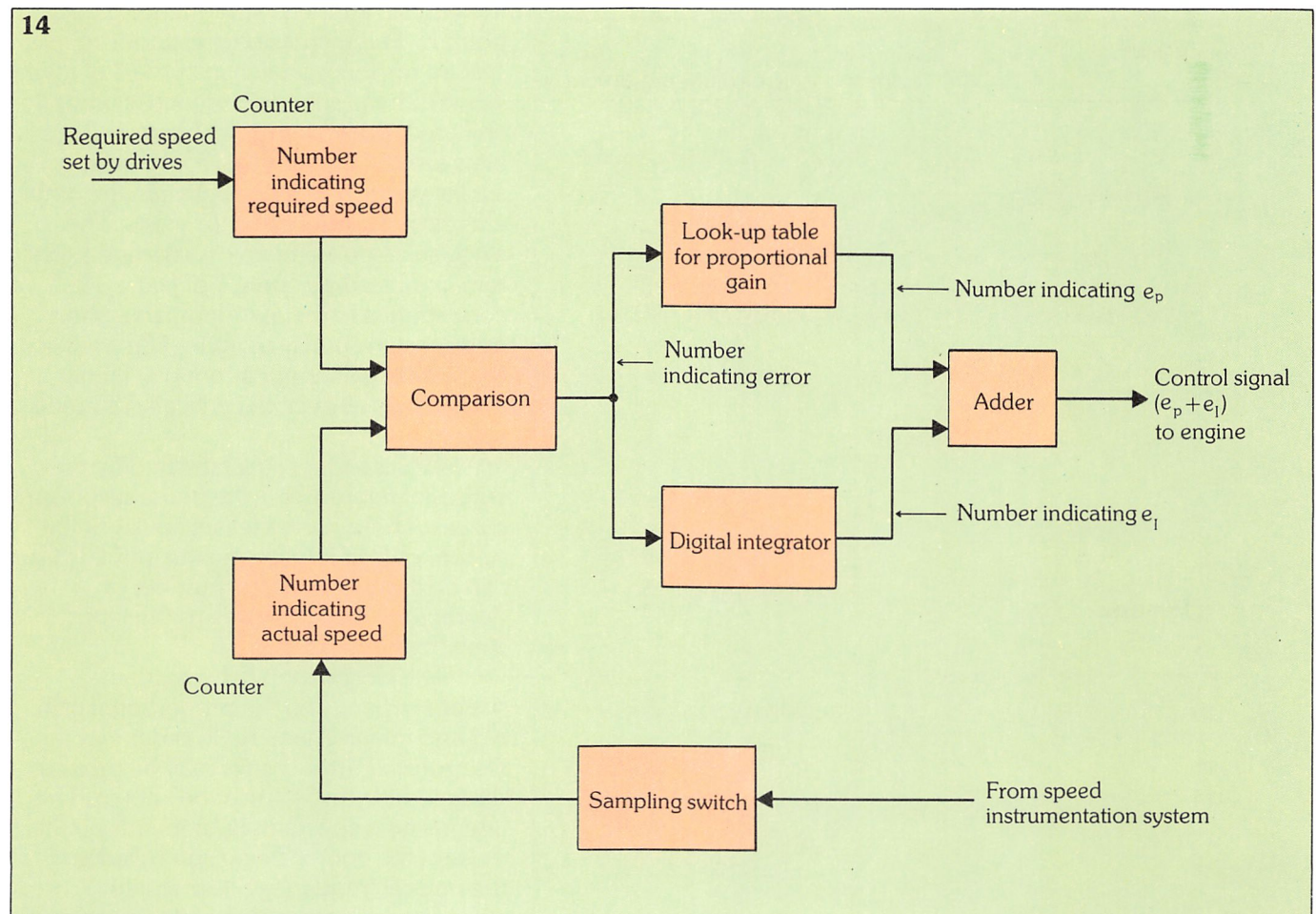
numbers over time and outputs a number representing integral gain output; e_I.

An adder sums the two numbers, producing the final output which is used to drive the throttle actuator, thus controlling car speed.

Such a PI cruise control electronics block may be constructed with a micro-computer – perhaps the same micro-computer used in our example trip computer instrumentation system – or may, alternatively, be built using analogue integrated circuits.

An analogue PI cruise control system circuit could be built, for example, using operational amplifiers.

Glossary

closed loop control	a control system in which feedback is used to form a complete loop in which an error signal is formed
closed loop system gain	overall gain of a feedback control system
cruise control	a control system used in a car to maintain constant speeds, demanded by the driver
damping coefficient	measure of the extent to which oscillations are reduced in a potentially oscillatory system
feedforward control	control system procedure, in which the effects of disturbances are calculated and allowed for, prior to the effect taking place
flowmeter	a transducer designed to produce an output signal proportional to the amount of liquid flowing through it
hunting	continuous oscillation of a control system around its required output
integral control	control procedure in which the control signal is proportional to an integral of the input or error signal
natural frequency, ω_n	the frequency at which a control system would oscillate if no damping was applied
open loop control	a control system procedure in which the system output is a direct consequence of the input and any disturbances
proportional control	control procedure in which the control signal is directly proportional to the input or error signal
proportional gain, K_p	gain of the proportional control block of a proportional control system
proportional-integral control (PI control)	control procedure where the control signal is obtained by adding the outputs of a proportional controller and an integral controller
steady state error	standing error (i.e. difference in actual output signal from required output) in a proportional control system
trip computer	car instrumentation system, generally allowing readings of car speed, fuel consumption, distance travelled etc.

The communications spectrum
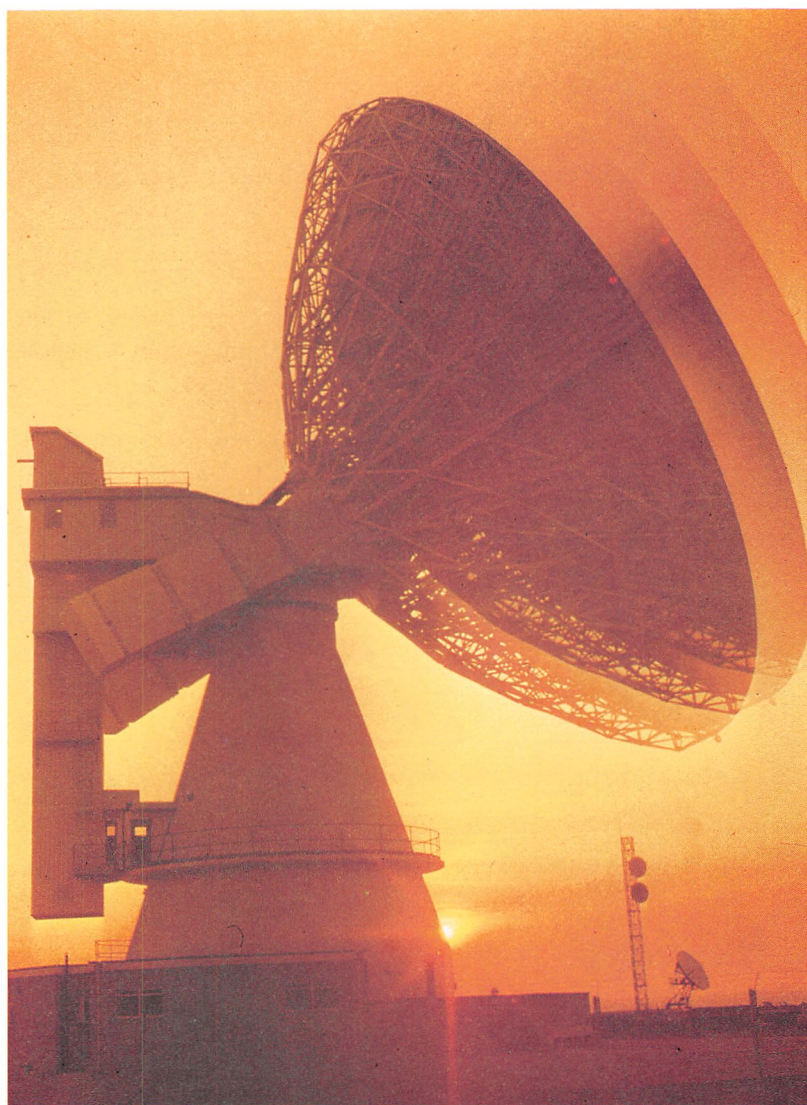
The electromagnetic spectrum

Telecommunications, literally meaning 'distance communications' is the branch of applied science that now encompasses all electrical and electronic methods of audio-visual communications and data transmission.

Fast, reliable communications links now form an integral part of both government and business practice, where facsimile transmission and modem links for computers mean that large amounts of data can be transmitted from one place to another. This series of articles will examine the different types of electronic communications device, and the methods and systems available.

Electromagnetic radiation is a form of energy which travels in waves, radiating away from their point of origination: such radiant energy is thus transmitted without physical contact.

Electromagnetic radiation has been defined by the scientist Maxwell as the energy radiated by a charged particle undergoing acceleration. This resulted in the generation of electric and magnetic fields which were mutually perpendicular and also perpendicular to the direction of motion of the radiation – termed **transverse electromagnetic fields**. If the charged particle moves sinusoidally, the instantaneous values of these fields are also sinusoidal and related to the charge density. Further work found that the behaviour of these electric and magnetic fields was identical to the behaviour of light, and that light itself was a form of electromagnetic radiation.

Maxwell found that the wavelength of an electromagnetic wave, when multiplied by its frequency, gives a figure which is equal to the velocity of light, and hence the

velocity of the wave, 3×10^8 ms^{-1} (*figure 1*). A graph for determining the wavelength from frequency is shown in *figure 2*.

Part of the electromagnetic spectrum is shown in *figure 3* and it can be seen that the range of frequencies used for communications purposes is considerably broader than that for visible light. Modern communications systems now use the

The Research House/British Telecom

Above: a receiving dish.

1. Sinusoidal electromagnetic wave.
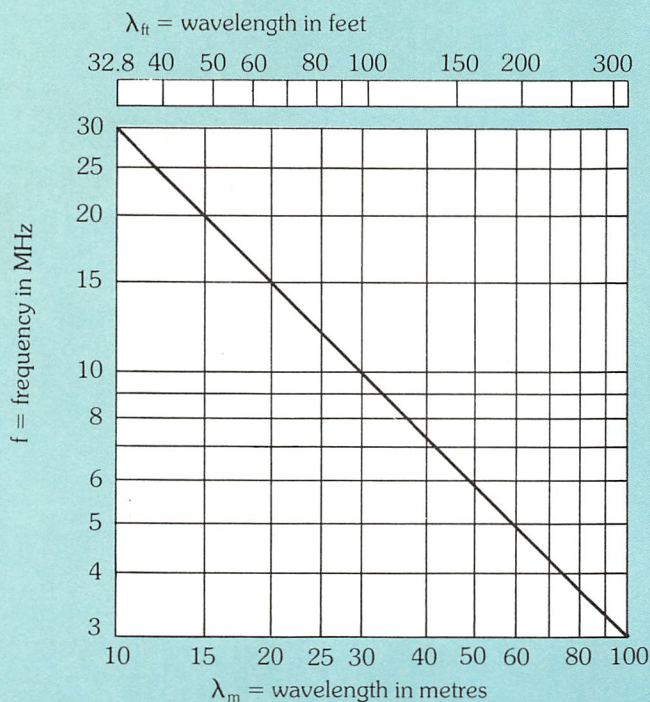

One wavelength
λ

A   A
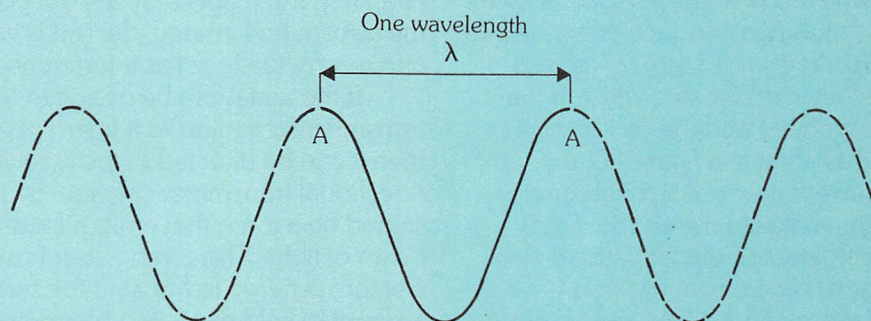
Sinusoidal electromagnetic wave, travelling through space at the speed of light, travels one wavelength during the time of one cycle.

$$f\,\lambda = 3 \times 10^8\,\text{ms}^{-1}\text{ (speed of light)}$$

$$\lambda = \frac{3 \times 10^8}{f}$$

2. Determining wavelength from frequency.

λ_{ft} = wavelength in feet

32.8 40 50 60 80 100 150 200 300

f = frequency in MHz

λ_m = wavelength in metres

For frequencies from		multiply f by	multiply λ_m by
0.03 –	0.3 MHz	0.01	100
0.3 –	3.0 MHz	0.1	10
3.0 –	30 MHz	1.0	1
30 –	300 MHz	10	0.1
300 –	3000 MHz	100	0.01
3000 –	30,000 MHz	1000	0.001
30,000 –	300,000 MHz	10,000	0.0001

range of frequencies from a few hertz up to, and including, the visible light range.

The sound spectrum
Sound is a form of energy which is transmitted in waves, however, unlike electromagnetic waves which can be propagated in a vacuum, sound waves depend on some medium for their transmission.

Healthy human ears are capable of perceiving sound over the range 20 Hz to 20 kHz; the ability to hear very high frequency sounds, however, deteriorates with age (an upper level of 10 kHz is considered to be average). The relationship between frequency, the sounds of familiar musical instruments and the notes of a piano is shown in *figure 4*.

The value obtained when multiplying together the wavelength and frequency of sound is $3.314 \times 10^2\,\text{ms}^{-1}$ – the velocity of sound in dry air. This value is much slower than the corresponding speed of electromagnetic waves. The electrical currents and electromagnetic radiation that exist at the same frequencies as sound, are used in electronic communications systems – their corresponding wavelengths, however, are about a million times longer than those of sound.

Generating electromagnetic radiation
If electrical currents of sufficient amplitude flow into an **aerial** (or antenna), then

795

electromagnetic radiation is produced. To examine why this is so, take a look at *figure 5a* which shows two conductors connected to a battery. As you can see, an electric field exists between the two wires. If the battery is removed and the conductors connected as shown in *figure 5b*, the electron current flow sets up a magnetic field at right angles to the electric field. (Remember, any conductor with an electric current passing through it is surrounded by a magnetic field and an electric field exists in any system in which there are equal positive and negative changes.) The important points to note here are: the magnetic field alternates with the electric field; both fields exist at right angles to each other; and the two fields are independent of each other.

Now, if the battery is replaced by an alternating current source (*figure 6*), then the result is a **dipole radiator** or **transmitting aerial**. The constantly alternating current sets up corresponding alternating electric and magnetic fields, creating a transmission of electromagnetic waves.

Figure 5 shows the state of the electric field at a particular polarity, but this, of course, changes with the alternating current. In fact, electromagnetic radiation can also be generated by devices, television receivers and motorised equipment for example, that are not supposed to produce it. This unwanted electromagnetic radiation can cause interference with the radiation that is deliberately generated to carry information. When electromagnetic radiation is generated by a radio transmitter, the pattern of energy transmitted depends mainly on the design of the aerial.
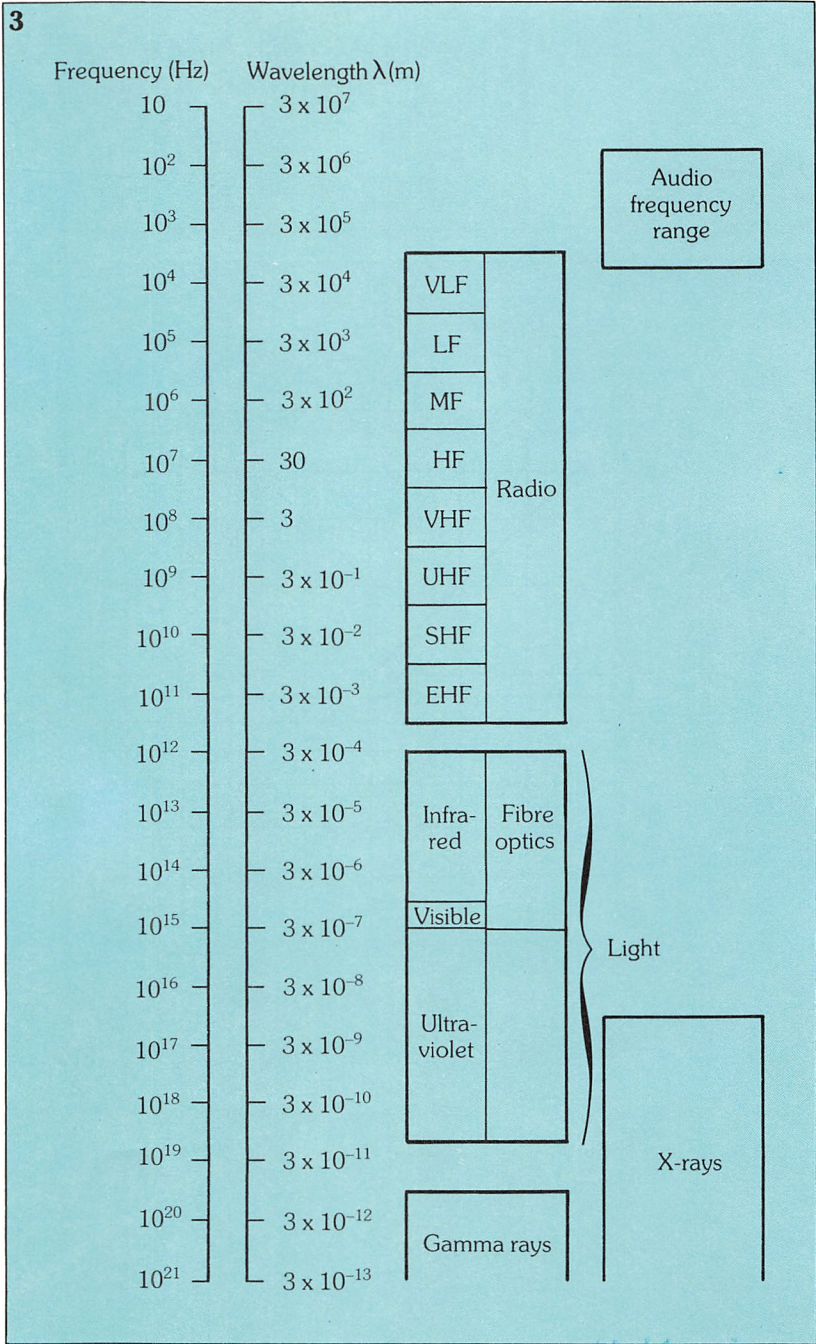
Radiation patterns
The simplest pattern of distribution of radiated energy is uniform or **omnidirectional**, where the same amount of energy is radiated in all directions. The more powerful the transmitter, the further this pattern extends into space; a conceptual analogy is that of a light bulb. This type of radiation pattern is useful if the receiver is to be anywhere in the space around the transmitter.

However, if the receiver is located in a fixed place, this transmitter arrangement is undesirable as *figure 7* shows. In this example, the receiver, a photodetector and amplifier, intercepts only a small portion of the transmitted energy, the rest is wasted and may actually act as interference.

If the aerial can be designed to transmit information as a beam of energy, then it can be directed to the receiver. The directional transmitter can now be represented by a torch that emits a focused beam of light. This arrangement uses less electrical energy to transmit the same amount of light onto the photodetector

3. Part of the electromagnetic spectrum showing the range of frequencies used for communications.

3

Frequency (Hz)	Wavelength λ (m)		
10	3×10^7		
10^2	3×10^6		Audio frequency range
10^3	3×10^5		
10^4	3×10^4	VLF	
10^5	3×10^3	LF	
10^6	3×10^2	MF	
10^7	30	HF	Radio
10^8	3	VHF	
10^9	3×10^{-1}	UHF	
10^{10}	3×10^{-2}	SHF	
10^{11}	3×10^{-3}	EHF	
10^{12}	3×10^{-4}		
10^{13}	3×10^{-5}	Infra-red	Fibre optics
10^{14}	3×10^{-6}		
10^{15}	3×10^{-7}	Visible	Light
10^{16}	3×10^{-8}		
10^{17}	3×10^{-9}	Ultra-violet	
10^{18}	3×10^{-10}		
10^{19}	3×10^{-11}		X-rays
10^{20}	3×10^{-12}		
10^{21}	3×10^{-13}	Gamma rays	

4. Relationship between frequency and the sounds of various musical instruments.
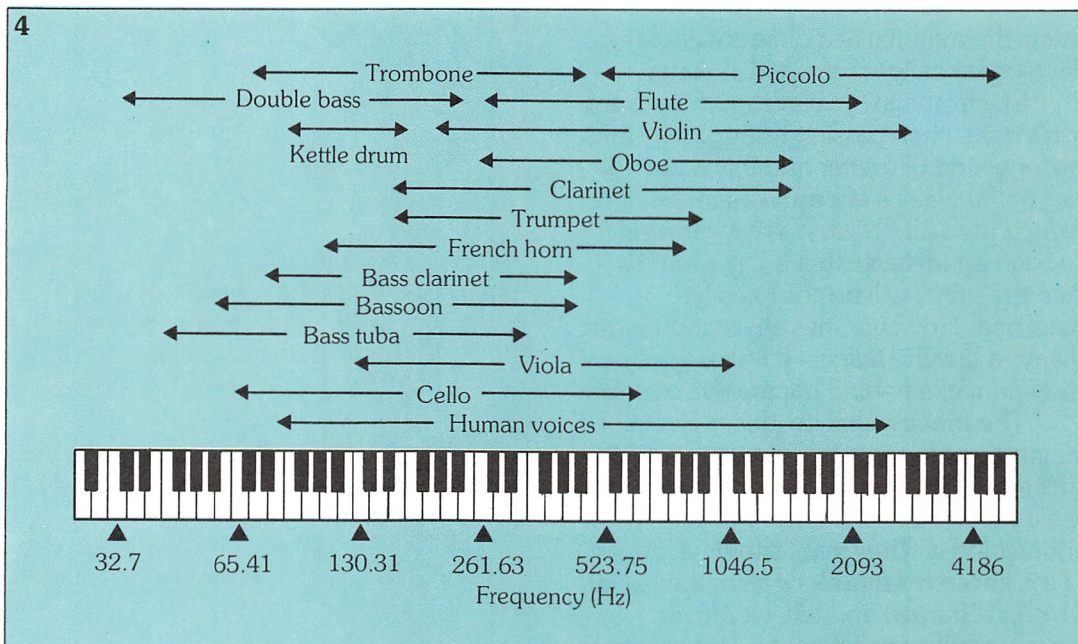
5. (a) An electric field between two wires connected to a battery; **(b)** the magnetic field is at right angles to the electric field.
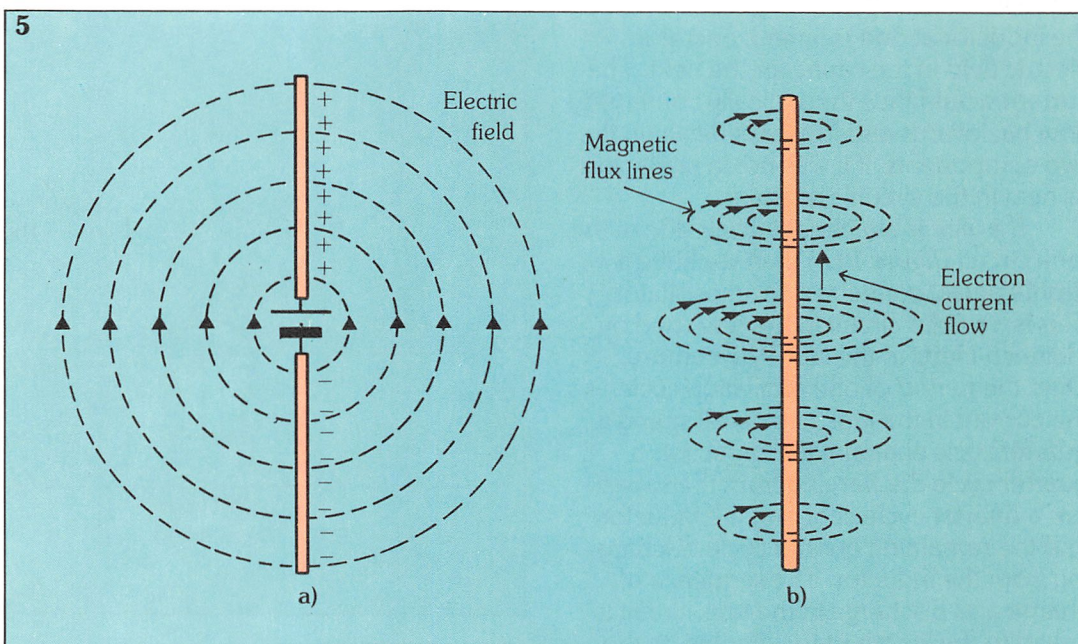
(*figure 8*).

As *figure 9* shows, this kind of beam irradiated pattern allows several similar transmitter/receiver pairs to operate in the same area without interfering with one another. One can think of the radiation as being channelled along 'pipes' of air, so that each communication system is entirely isolated from all others. Again, the stronger the transmitter beam, the greater the distance can be between transmitter and receiver. One factor to be taken into consideration is that electromagnetic radia-

tion can be affected by the atmosphere through which it is travelling, the amount by which the strength is reduced depending on many different factors.

Electromagnetic waves can be reflected by electrically conductive surfaces, like metals, the angle of reflection being equal to the angle of incidence (*figure 10a*). If electromagnetic waves meet nonconducting surfaces – for example plastics – the signal will be partially reflected and partially absorbed (*figure 10b*). Notice that the portion of the signal that is absorbed
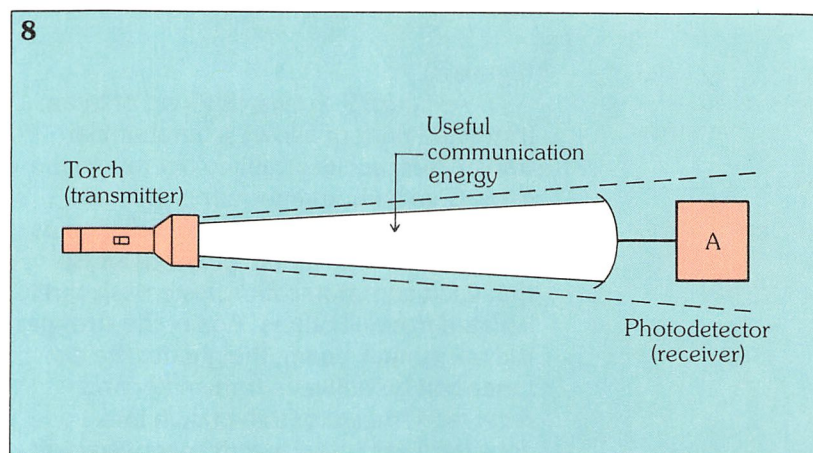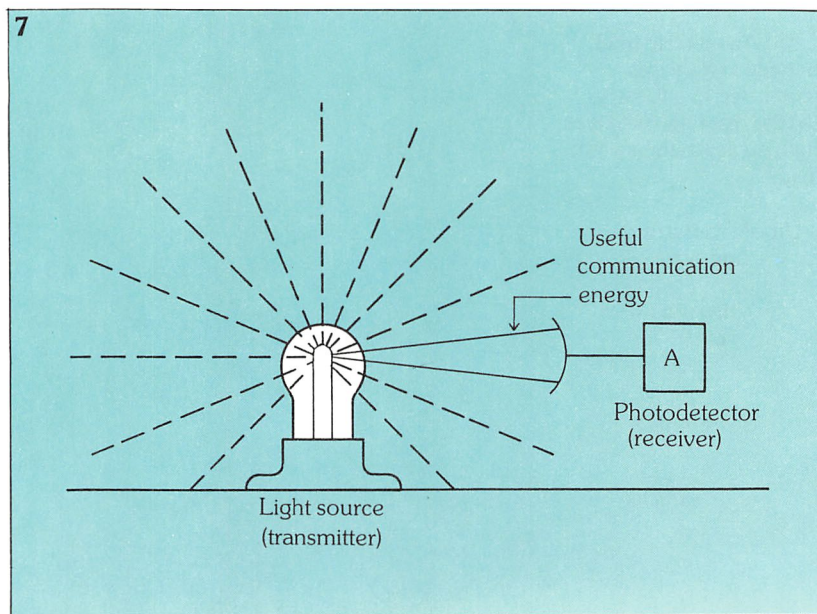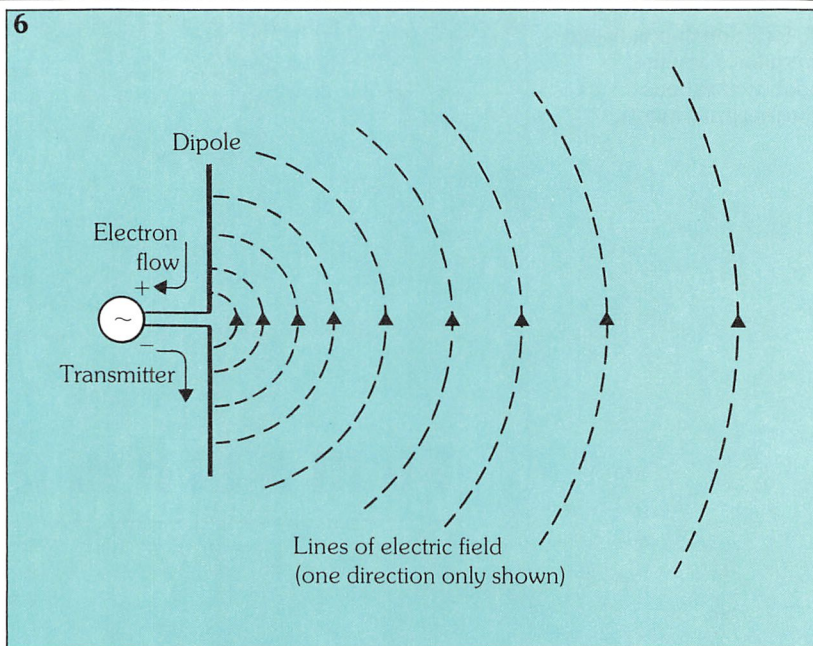
enters the material at a different angle — like a beam of light refracted in water.
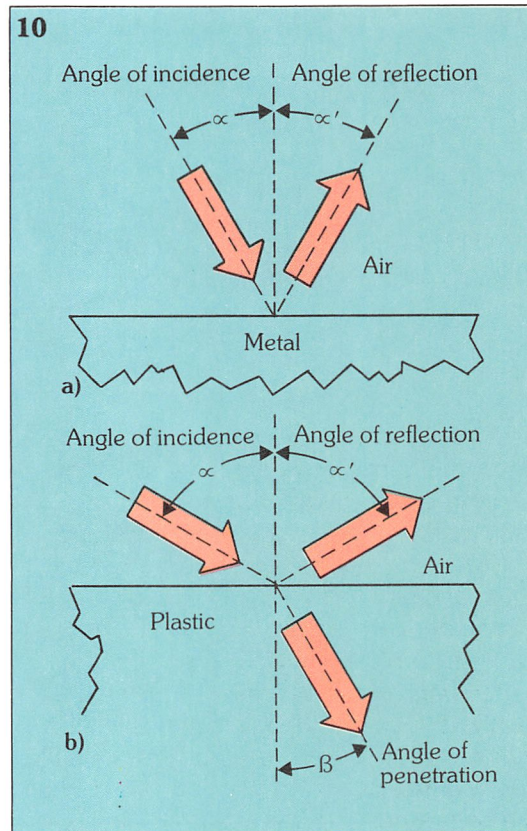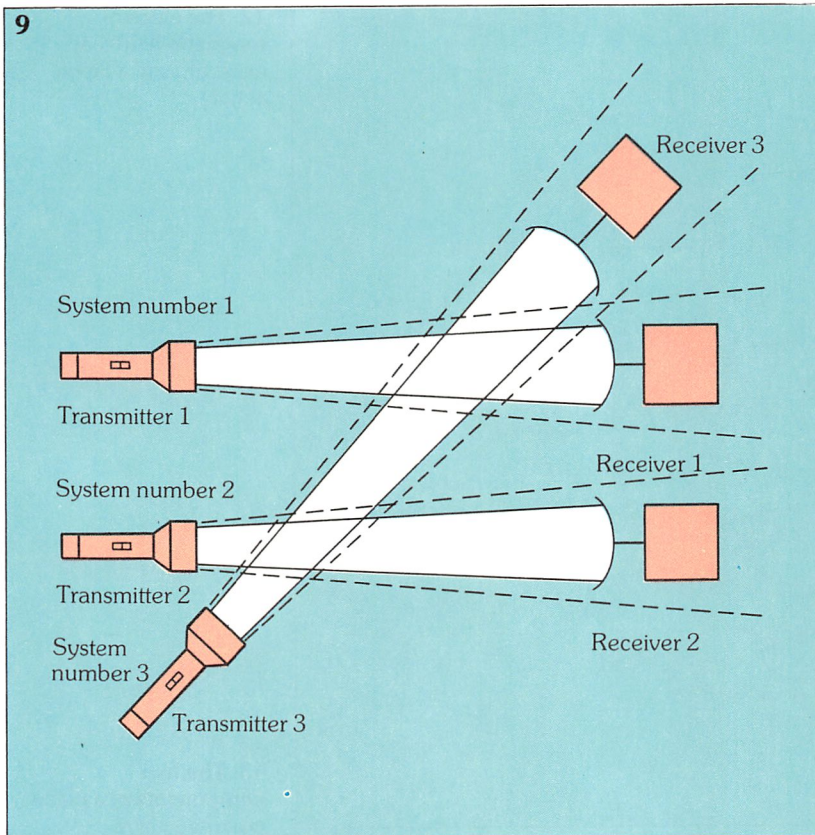
Electromagnetic waves suffer absorption losses when passing through, or striking, any kind of matter and this is due to part of the wave's energy content being transformed into heat. If a wave travels through an obstacle that is big enough, then the signal will be completely absorbed. However, this absorption is not always a disadvantage — it forms the operating principle behind microwave ovens.

The process that produces electromagnetic waves is very complex and advanced mathematics is needed to describe it; a qualitative explanation will help to understand it. The circuit shown in *figure 11a* is known as a **tank** or resonant circuit, which, if charged and left, would, in theory, oscillate indefinitely. The energy in the circuit is stored as a magnetic field in the inductor at one moment, and as an electric field in the capacitor the next. The current would thus theoretically continually flow backwards and forwards between the two components, if it was not in reality lost as heat in the circuit's resistance.

If a wire is attached to each side of the tank circuit (*figure 11b*) then we have a **dipole aerial** connected to an oscillator. This is a similar circuit to the one used by Heinrich Hertz in the late 19th century. Over the period of one oscillation cycle, the current in the tank circuit will spend a quarter cycle charging the capacitor; a quarter cycle discharging from the capacitor; a quarter cycle charging the inductor; and the remaining quarter cycle discharging from the inductor. The sequence of charge and discharge in the tank circuit is linked to the length of the dipoles, as the electron current travels to the end of each wire and back to the tank circuit creating alternate electric and magnetic fields. The length of the wires has to be such that the electron current can complete a circuit of both dipoles in the time taken for the oscillator to complete one cycle, otherwise the oscillations in the aerial will not be in phase with the oscillations in the tank circuit. The length of each dipole is equal to a quarter wavelength; both dipoles together being equal to a half-wave. As:

$$\text{wavelength, } \lambda = \frac{\text{speed of light}}{\text{frequency}}$$

Lines of electric field (one direction only shown)

Light source (transmitter)

9

Receiver 3

System number 1

Transmitter 1

System number 2

Transmitter 2

System number 3

Transmitter 3

Receiver 1

Receiver 2

10

Angle of incidence | Angle of reflection

α α'

Air

Metal

a)

Angle of incidence | Angle of reflection

α α'

Air

Plastic

b)

β Angle of penetration

**6. Dipole radiator** or transmitting aerial.

7. Omnidirectional radiation pattern: a fixed receiver intercepts only a small portion of the transmitted energy.

8. Radiation pattern in the form of a directed beam.

9. Several beam irradiated patterns may operate in the same area without interference.

10. Electromagnetic waves reflected by: **(a)** metals: and partially absorbed by **(b)** non-conducting surfaces.

11. (a) Tank; (b) dipole aerial connected to an oscillator.

11

$\frac{\lambda}{4}$

$\frac{\lambda}{2}$

a)

b)
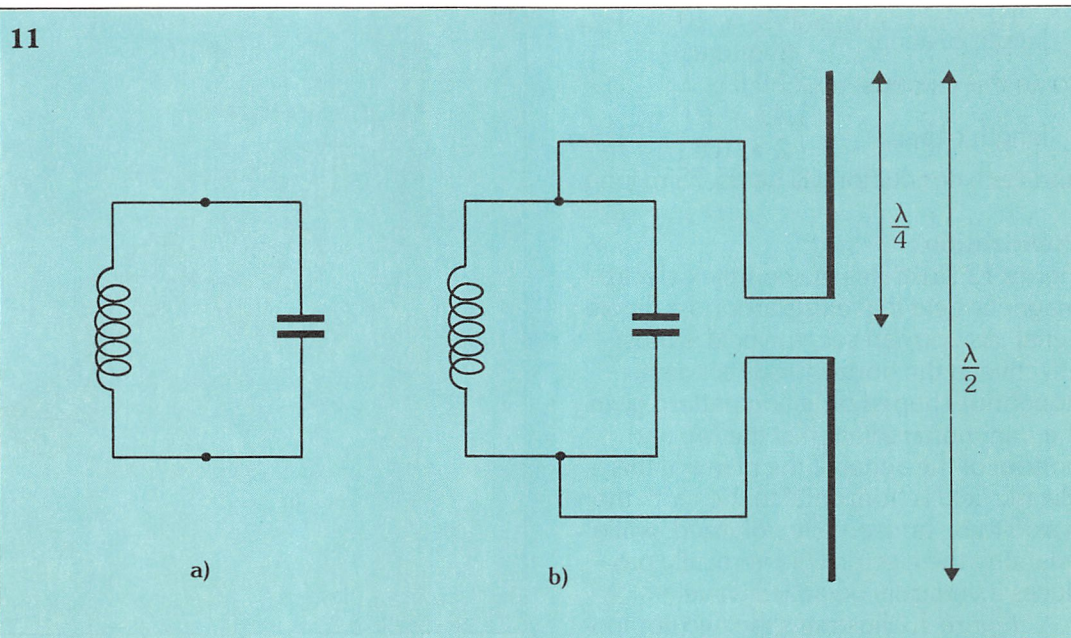
we can determine the dimensions of the dipole needed to transmit a signal of, say, 2 MHz:
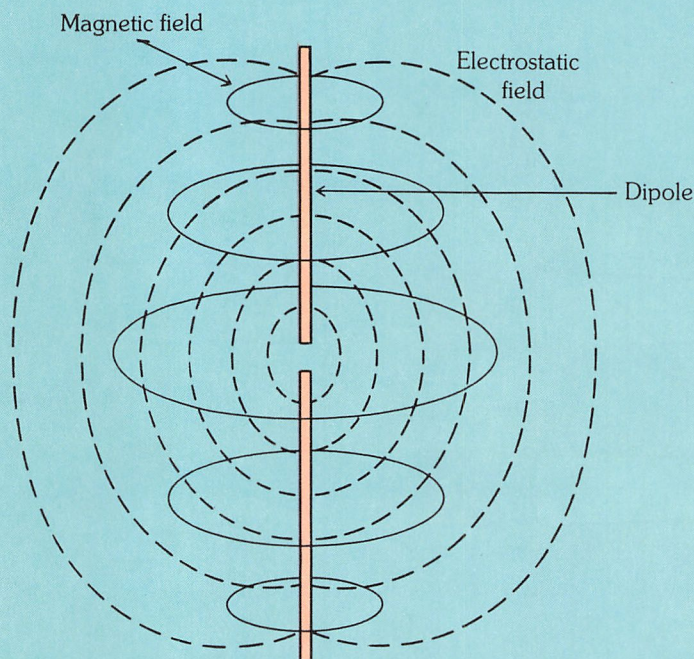
$$\text{wavelength,} \quad \lambda = \frac{3 \times 10^8}{2 \times 10^6} = 150 \text{ m}$$

A half-wavelength is therefore 75 m long, so each conductor in the dipole will be

37.5 m long.

However, electromagnetic waves cannot travel as quickly in conductors (having resistance) as they do in space, so this value needs to be corrected. The formula for the length of dipole aerials used to carry signals of up to 30 MHz is:

Magnetic field

Electrostatic field

Dipole

12. The electro-magnetic field that exists around a dipole aerial.

13. The two components of a plane polarized wave.

$\frac{\lambda}{4}$ Dipole

Electric field

$\frac{\lambda}{4}$ Magnetic field

Direction of propagation

$$\text{length of aerial} = \frac{1.43 \times 10^8}{\text{frequency}}$$

so, in this example (at 2 MHz):

$$\text{length of aerial} = \frac{1.43 \times 10^8}{2 \times 10^6} = 71.5 \text{ m}$$

and each conductor will be 35.75 m long.

Polarization

Figure 12 illustrates the complex electro-magnetic field that exists around a dipole aerial. As you can see, the field is most effective in the horizontal direction: its doughnut-shaped radiation pattern giving the minimum radiation at the top and bottom of the aerial. If the plane of the electric field is horizontal to the earth then it is said to be horizontally polarized, while orienting the electric field vertically pro-duces a vertically polarized wave.

Figure 13 illustrates the two compo-nents of a vertically polarized electro-magnetic wave — whether horizontal or vertical this is known as a **plane polarized wave**. When a dipole aerial is vertical to the earth, its doughnut-shaped radiation pattern will be horizontal, and the max-imum radiation will therefore be in all horizontal directions around the aerial. If,
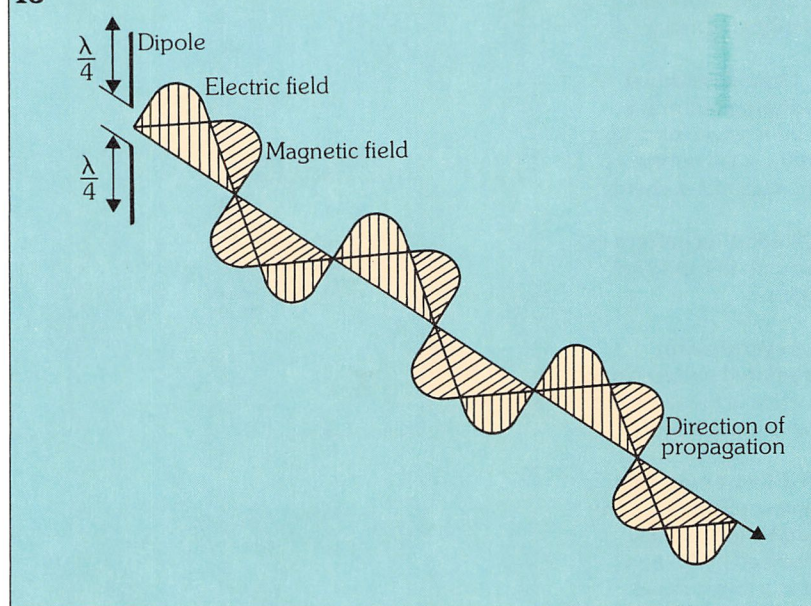
on the other hand, the aerial is horizontal to the earth, then the doughnut pattern will be vertical and the maximum radiation will be in this vertical plane. This means that a dipole aerial may be used to transmit electromagnetic waves in a directional manner. *(continued in part 26)*